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

Volume 5 Issue 6

June 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 June 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 6 June 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

An Open Source P2P Encrypted Voip Application

Ajay Kulkarni
Operations & Cross Product Technology
Barclays Investment Bank
Pune, India

Saurabh Kulkarni
Software Engineer
Accenture
Mumbai, India

***Abstract*—Open source is the future of technology. This community is growing by the day; developing and improving existing frameworks and software for free. Open source replacements are coming up for almost all proprietary software nowadays. This paper proposes an open source application which could replace Skype, a popular VoIP soft phone. The performance features of the developed software is analyzed and compared with Skype so that we can conclude that it can be an efficient replacement. This application is developed in pure Java using various APIs and package and boasts features like voice calling, chatting, file sharing etc. The target audience for this software will initially only be organizations (for internal communication) and later will be released on a larger scale.**

Keywords—voip; softphone; java; open source

I. INTRODUCTION

Email was the original killer application for the Internet. Today, voice over IP (VoIP) and instant messaging (IM) are fast supplementing email in both enterprise and home networks. Skype is an application that provides these VoIP and IM services in an easy to use package that works behind Network Address Translators (NAT) and firewalls. It has attracted a user-base of 70 million users, and is considered valuable enough that Microsoft recently acquired it for $8.5 billion [1]. In this paper, we present an open source replacement for Skype and a measurement study between the developed Peer-to-Peer application and Skype. While measurement studies of both P2P file sharing networks [2] and traditional VoIP systems [3] have been performed in the past, little is known about VoIP systems that are built using a P2Parchitecture.

One of our key goals in this paper is to understand how efficient this P2P VoIP application is to replace a giant like Skype which is also a P2P application. A peer-to-peer VoIP network typically consists of a core proxy network and a set of clients that connect to the edge of this proxy network (Fig. 6). This network allows a client to dynamically connect to any proxy in the network and to place voice calls to other clients on the network. VoIP uses the two main protocols: route setup protocol for call setup and termination, and Real-time Transport Protocol (RTP) [9] for media delivery. In order to satisfy QoS requirements, a common solution used in peer-to-peer VoIP networks is to use a route setup protocol that sets up the shortest route on the VoIP network from a caller source to a receiver destination. RTP is used to carry voice traffic between the caller and the receiver along an established bi-directional voice circuit.

Now, talking about the software license, as the developed software is open source its source code is freely available to all for further development. Its free availability gives scope for peer review, regular bug fixes and hence there is an increase in reliability of the application. Security flaws can be analyzed by anyone and can be fixed as and when a loophole is discovered. These are just some of the key points on why open source is preferred over proprietary software these days, the full list is endless.

Overall, this paper makes three contributions. First, light is shed on the VoIP network construction. Second, the architecture and design of the developed software is described in detail. Third, a comparison is done between the developed P2P VoIP application and Skype.

II. RELATED WORK

Skype offers three services: VoIP allows two Skype users to establish two-way audio streams with each other and supports conferences, IM allows two or more Skype users to exchange small text messages in real-time, and file-transfer allows a Skype user to send a file to another Skype user (if the recipient agrees). Skype also offers paid services that allow Skype users to initiate and receive calls via regular telephone numbers through VoIP-PSTN gateways.

Despite its popularity, little is known about Skype's encrypted protocols and proprietary network. Skype is related to KaZaA; both the companies were founded by the same individuals, there is an overlap of technical staff, and that much of the technology in Skype was originally developed for KaZaA. Network packet level analysis of KaZaA [14] and of Skype [15] support this claim by uncovering striking similarities in their connection setup, and their use of a "supernode"-based hierarchical peer-to-peer network.

Supernode-based peer-to-peer networks organize participants into two layers: supernodes, and ordinary nodes. Such networks have been the subject of recent research in [16]. Typically, supernodes maintain an overlay network among themselves, while ordinary nodes pick one (or a small number of) supernodes to associate with.

III. VOIP OVERVIEW

The section below describes the working of VoIP networks based on the function of the network components listed in Figure 7. Depending upon the particular network architecture [4] some of these network components [6] may be combined into a single solution.

A. Call Agent/Sip Server/Sip Client

The Call Agent/SIP Server/SIP Client is located in the service provider's network and provides call logic and call control functions, typically maintaining call state for every call in the network. The Call Agent will participate in signaling and device control, terminating or forwarding messages. There are numerous relevant protocols depending upon the network architecture including SIP (Session Initiation Protocol), SIP-T, H.323, BICC, H.248, MGCP/NCS, SS7, AIN, ISDN, etc. [19, 21]. A SIP Server provides equivalent function to a Call Agent in a SIP signaling network, its primary roles are to route and forward SIP requests, enforce policy (for example call admission control) and maintain call details records. For example the SIP Server in Service Provider 1's network will route and forward SIP requests from SIP Phones belonging to customers. A SIP Client provides similar function to a SIP Server, but originates or terminates SIP signaling rather than forwarding it to a SIP Phone or other Customer Premises Equipment. The Call Agent/SIP Server terminates the SIP signaling and converts it to H.248 or MGCP to set up a call to the correct subscriber. Call Agents are also known as Media Gateway Controllers, Soft switches and Call Controllers. All these terms convey a slightly different emphasis but maintaining call state is the common function reused with other services and to create new value added services.

B. Service Broker

The service broker is located on the edge of the service provider's service network and provides the service distribution, coordination, and control between application servers, media servers, call agents, and services that may exist on alternate technologies (i.e. Parlay Gateways and SCP's). The service broker allows a consistent repeatable approach for controlling applications in conjunction with their service data and media resources to enable services, to allow services to be reused with other services and to create new value added services.

C. Application Server

The Application Server is located in the service provider's network and provides the service logic and execution. Typically the Call Agent will route calls to the appropriate application server when a service is invoked that the Call Agent cannot support itself.

D. Media Server

This Media Server is located in the service provider's network. It is also referred to as an announcement server. For voice services, it uses a control protocol, such as H.248 or MGCP, under the control of the call agent or application server. Some of the functions the Media Server can provide are codec transcoding and voice activity detection, tone detection and generation and interactive voice response (IVR) processing.

E. Signalling Gateway

The Signaling Gateway is located in the service provider's network and acts as a gateway between the call agent signaling and the SS7-based PSTN. It can also be used as a signaling gateway between different packet-based carrier domains. It provides signaling translation, for e.g. between SIP and SS7 (Signaling System 7) or simply signaling transport conversion e.g. SS7 over IP to SS7 over TDM.

F. Trunking Gateway

The Trunking Gateway is located in the service provider's network and as a gateway between the carrier IP network and the TDM (Time Division Multiplexing)-based PSTN. It provides transcoding from the packet-based voice, VoIP onto a TDM network. Typically, it is under the control of the Call Agent / Media Gateway Controller (MGC) through a device control protocol such as H.248 or MGCP.

G. Access Gateway

The Access Gateway is located in the service provider's network. It provides support for POTS phones and typically, it is under the control of the Call Agent / Media Gateway Controller through a device control protocol such as H.248 or MGCP.

H. Bandwidth Manager

The Bandwidth Manager is located in the service provider's network and is responsible for providing the required QoS from the network. It is responsible for the setting up and tearing down of bandwidth within the network and for controlling the access of individual calls to this bandwidth.

I. Bridge/Router

The Bridge/Router is located at the customer premises and terminates the WAN (Wide Area Network) link at the customer premises. Voice services for example SIP phones, can be bridged/routed via this device.

J. IP Phone/Microphone

IP Phones and Microphones are located at customer premises and provide voice services. They interact with the Call Agent/SIP Server using a signaling protocol such as SIP, H.323 or a device control protocol such as H.248 or MGCP.

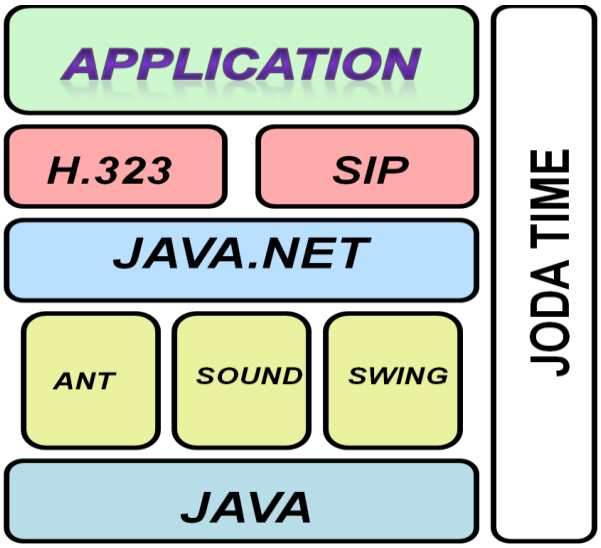


Fig. 1. Architecture of the developed software

IV. The VoIP Open Source Project

The open source software "movement" has received enormous attention in the last several years [11]. It is often characterized as a fundamentally new way to develop software [7] that poses a serious challenge [8] to the commercial software businesses that dominate most software markets today. The challenge is not the sort posed by a new competitor that operates according to the same rules but threatens to do it faster, better, cheaper. The developed open source application is developed as a replacement for proprietary VoIP software. The network architecture (Fig. 5) for the designed software is really simple to implement. Various Java APIs and packages like Swing, Java Sound API, Java.net package, Joda-Time API are used for the implementation of this project (Fig. 1). The Video Call feature is under development and will be released in version 2.0 of the software; features already implemented include P2P chatting, file sharing & encrypted voice call. Let us understand the implementation and function of every feature in the developed open source application. An extra feature of server monitoring is added for the organization centric release of this open source software.

A. Database

The MySQL database is used to store all user data. The password is encrypted and stored along with the username (primary key); the profile picture of the user is stored in the form of BLOB (Binary Large Object) in the database.

B. Login/Sign Up:

The login page in Fig. 2 is a simple form with two fields for username and password; when sign in is clicked a query runs on the database to check the validity of the credentials. The sign up page takes all necessary information from the user and creates an account for the user by storing his data securely in the database

C. Home

All user functions are displayed on the home page in Fig. 3. The user can update his profile picture, check online users, start a chat session, and make a voice call and even share files with another user. To end the session, the user can click on sign out.

D. Chat Box

A user can chat with multiple users at a time. The chat is implemented in the Java.net package and the delay is message exchange is virtually zero. The chat window is the big white box in Fig. 3.

E. Encrypted Voice Call

The Java Sound API is the backbone of this feature giving all the necessary support to it. A custom voice call package has also been developed to encrypt voice data packets (using a custom encryption algorithm) [10, 21], improve sound quality during the call and also to minimize delay and the echo effect.

F. File Share

Users can exchange files by clicking on the button on the left bottom of the home screen in Fig. 3, of any format except .exe between each other. The file sharing module is implemented using Java.net using simple TCP/IP port programming. The application has been tested for file sizes up to 20 MB and delay observed is negligible. The file is stored on the server for just a fraction of a second to prevent server overload.

G. Chat Monitoring

This is not a P2P feature but it is specially included for the organization release of the software. The chat monitoring window helps organizations to keep a track of conversations between its employees for compliance purposes.

H. Hardware

For the initial testing of the software, the application and database server was a remote computer with basic configurations. The client is really light and is currently supported only on Windows. This just shows how light and efficient open source applications are and why they are getting more and more popular every day.

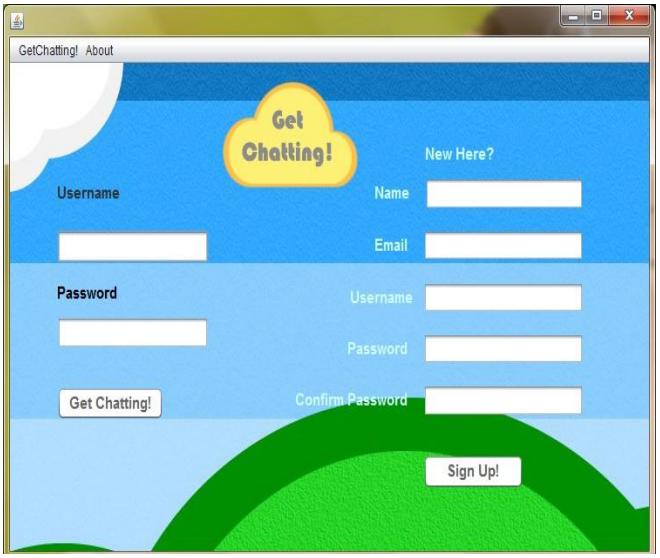


Fig. 2. Login screen of software

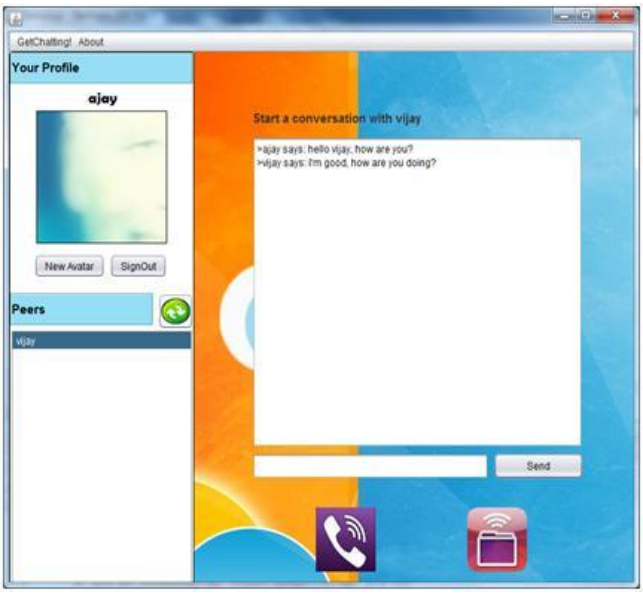


Fig. 3. Home screen of software

V. Open Source Application V/S Skype

Both the applications were tested on a network bandwidth of 2 Mbps and on a system with basic configurations. Experimental result [5, 12] for every feature is as follows:

A. Voice Calling

The quality [18] of calling in both applications were analysed based on the clarity and delay in transmission of voice from source to destination and the results are presented in Table I.

TABLE I. Voice Quality of Skype Vs. OS Application

Application	Clarity	Delay
Skype	High	22 ms
Open Source App	High	31 ms

B. File Sharing

A standard file size of 2.5 Mb was used to test the results of file transfer. The upload and download speed is shown in Table II.

TABLE II. File Share speed of Skype vs. OS application

Application	Upload Speed	Download Speed
Skype	16 sec	11 sec
Open Source	19 sec	13 sec

C. Chatting

P2P chatting is seamless in both applications with practically no delay. The organizational version of this feature in which the messages a routed through a monitoring server was also tested for delay and the results indicated that there is negligible delay as compared to the P2P version

D. Handling Bulky Files – Stress Test

Both applications were stress tested by sending three files of size 25 MB each, back to back, after every 5 seconds. While the Skype window froze for a while but was back on track and started transmitting data, the Open Source Application crashed while sending the third file.

Thus stability is an issue which needs to be addressed in further releases.

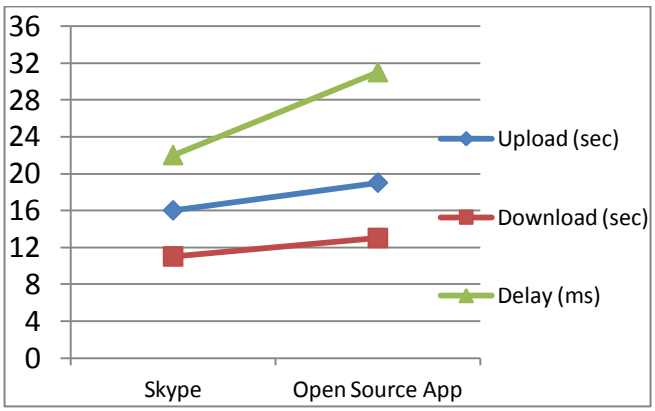


Fig. 4. Skype v/s Open source application performance

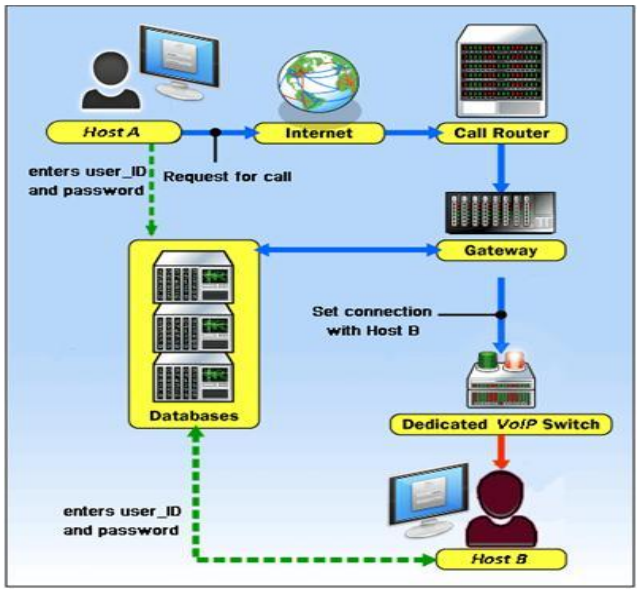


Fig. 5. System architecture of the open source application

VI. Future Works

The developed P2P open source application is currently being released for organizations and the full-fledged version for all users will be released in version 2.0. The Linux based version of the software is also in the works and will make its way out to user at the same time. Version 2.0 will also include features like Video Call [13] and Screen Share which will also be developed in Java. Users will also see an increase in call quality and numerous GUI tweaks. Also performance issues will be taken care of and random crashes due to large amounts of data transmission will be fixed in subsequent releases. The software will be licensed under the open source license and will be made public via a website which is also under development.

VII. Conclusion

This paper presents an open source VoIP application to replace proprietary software like Skype. From the experimental data is gathered it is evident that the open source application can perform as good as Skype. The time delay in call routing and voice data transfer of the developed software is minimum considering that its voice data packets are being encrypted, and its performance can match that of Skype (refer Fig. 4). File sharing speeds are matching that of the proprietary software and hence the open source application performs up to the mark. The present experimental results are just preliminary in nature and further study is required on this topic. The GUI of Skype is really user friendly and has matured over the years; this is one area where the developed software has to catch up to a considerable extent.

Overall, the measurement data presented is useful for designing and modeling a peer-to-peer VoIP system. The architecture of the open source application in Fig. 5 lays down a foundation for all future VoIP system designing activities. The open source application can be further developed and tweaked by the community and can hopefully one day replace proprietary VoIP software.

REFERENCES

[1] Andrew Sorkin and Steve Lohr, Microsoft to Buy Skype for $8.5 Billion. *The New York Times* (May 10, 2011).

[2] Pouwelse, J., Garbacki, P., Epema, D., and Sips, H, The bittorrent p2p file-sharing system: Measurements and analysis. In *Proceedings of the IPTPS '05* (Ithaca, NY, Feb. 2005).

[3] Calyam, P., Sridharan, M., Mandrawa, W., and Schopis, P. Performance measurement and analysis of h.323 traffic. In *Proceedings of the 5th International Workshop on Passive and Active Network Measurement (PAM 2004)* (Antibes Juan-les-Pins, France, Apr. 2004).

[4] Cisco System, Data Considerations and Evolution of Transmission Network Design, http://www.cisco.com/en/US/prod/collateral/optical/ps5724/ps2006/prod_white_paper0900aecd803faf8f_ps2001_Products_White_Paper.html, 2009.

[5] The Network Simulator—ns2, http://www.isi.edu/nsnam/ns/, 2007.

[6] Paul Drew, Chris Gallon, Next-Generation VoIP Network Architecture, *MSF Technical Report*, March 2003.

[7] C. DiBona, S. Ockman, and M. Stone, Open Sources: Voices from the Open Source Revolution. Sebastopol, CA: O'Reilly, 1999. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev.

[8] P. Vixie, "Software Engineering," in Open Sources: Voices from the Open Source Revolution, C. DiBona, S. Ockman, and M. Stone, Eds. Sebastopol, CA: O'Reilly, 1999, pp. 91-100. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[9] W. Mazurczyk and Z. Kotulski, "New Security and Control Protocol for VoIP Based on Steganography and Digital Watermarking," tech. rep., Institute of Fundamental Technological Research, Polish Academy of Sciences, June 2005, http://arxiv.org/ftp/cs/papers/0602/0602042.pdf

[10] D. Kundur and K. Ahsan, "Practical Internet Steganography: Data Hiding in IP," *Proc. Texas Wksp. Security of Information Systems*, Apr. 2003.

[11] Elliott M, Scacchi W. 2003. Free software developers as an occupational community: Resolving conflicts and fostering collaboration. In *Proceedings of the ACM International Conference on Supporting Group Work*, Sanibel Island, FL, November 2003, 21–30.

[12] Marc Greis, "Tutorial for Network Simulator NS", http://www.scribd.com/doc/13072517/tutorial-NS-full-byMARC-GREIS.

[13] Ajay Kulkarni, Saurabh Kulkarni, Ketki Haridas and Aniket More. Article: Proposed Video Encryption Algorithm v/s Other Existing Algorithms: A Comparative Study. International Journal of Computer Applications 65(1):1-5, March 2013. Published by Foundation of Computer Science, New York, USA

[14] Liang, J., Kumar, R., and Ross, K. W. The kazaa overlay: A measurement study. Computer Networks 49, 6 (Oct. 2005).

[15] Baset, S. A., and Schulzrinne, H. An Analysis of the Skype Peer-to Peer Internet Telephony Protocol. In Proceedings of the INFOCOM '06 (Barcelona, Spain, Apr. 2006).

[16] Xu, Z., and Hu, Y. Sbarc: a supernode based peer-to-peer file sharing system. In Proceedings of the 8th IEEE Symposium on Computers and Communications (ISCC'03) (Antalya, Turkey, July 2003).

[17] Castro, M., Costa, M., and Rowstron, A. Debunking some myths about structured and unstructured overlays. In *Proceedings of the NSDI '05* (Boston, MA, May 2005).

[18] S. Jadhav, H. Zhang, Z. Huang, - Performance Evaluation of Quality of VoIP in WiMAX and UMTS‖ PDCAT (2011), pp. 378

[19] S. Sahabudin, M.Y. Alias. End-to-end delay performance analysis of various codecs on VoIP Quality of Service. Communications (MICC), 2009 IEEE 9th Malaysia International Conference on. vol., no., pp.607-612, 15-17 Dec. 2009.

[20] Yan Zhang and Huimin Huang, (2011) "VOIP voice network technology security strategies", Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp 3591-3594

[21] S. Ghosh, "Comparative Study of QOS Parameters of SIP Protocol in 802.11a and 802.11b Network", International journal of Mobile Network Communications & Telematics (IJMNCT), 2 (6), 21-30. (2012).

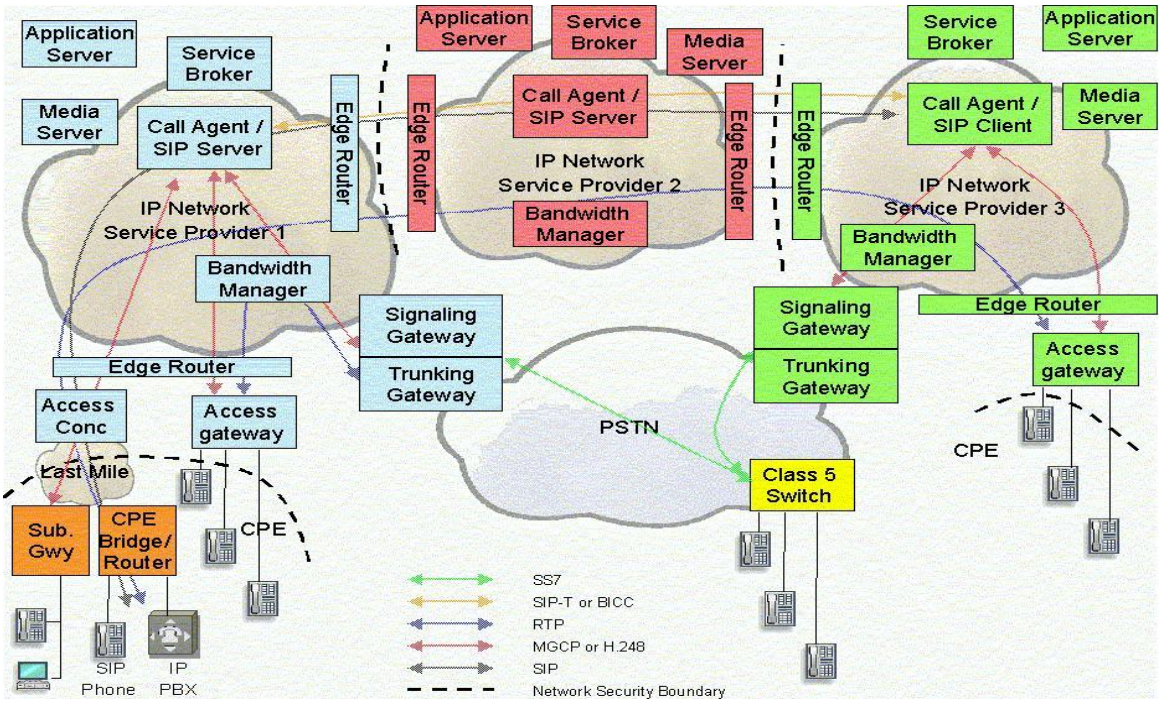


Fig. 6. Network architecture and components of a typical VoIP system

The Coverage Analysis for Low Earth Orbiting Satellites at Low Elevation

Shkelzen Cakaj[1], Bexhet Kamo[1], Algenti Lala[1], Alban Rakipi[1]
[1] Faculty of Information Technology
Polytechnic University of Tirana
Tirana, Albania

***Abstract*—Low Earth Orbit (LEO) satellites are used for public networking and for scientific purposes. Communication via satellite begins when the satellite is positioned in its orbital position. Ground stations can communicate with LEO satellites only when the satellite is in their visibility region. The duration of the visibility and the communication vary for each LEO satellite pass over the station, since LEO satellites move too fast over the Earth. The satellite coverage area is defined as a region of the Earth where the satellite is seen at a minimum predefined elevation angle. The satellite's coverage area on the Earth depends on orbital parameters. The communication under low elevation angles can be hindered by natural barriers. For safe communication and for savings within a link budget, the coverage under too low elevation is not always provided. LEO satellites organized in constellations act as a convenient network solution for real time global coverage. Global coverage model is in fact the complementary networking process of individual satellite's coverage. Satellite coverage strongly depends on elevation angle. To conclude about the coverage variation for low orbiting satellites at low elevation up to 10º, the simulation for attitudes from 600km to 1200km is presented through this paper.**

Keywords—LEO; satellite; coverage

I. Introduction

Generally, satellite' circular orbits are categorized as Geosynchronous Earth Orbits (GEO), Medium Earth Orbits (MEO) and Low Earth Orbits (LEO). The main difference among them is in the attitude above the Earth surface [1]-[3].

The satellites traversing in orbits of attitudes up to around 1400 km (limited by Van Allen belt [4]) are considered as LEO satellites. LEO satellites are moving at around 7.5 km/s velocity relative to a fixed point on the Earth (ground station) [5]. The characteristics of LEOs are: the *shortest distance* from the Earth compared with other orbits and consequently *less time delay*. These characteristics make them very attractive even for scientific applications or communications networking [5], [6].

The single satellite coverage area is defined as a region of the Earth where the satellite is seen at a minimum predefined elevation angle. For multi satellite coverage or global coverage the management policy for satellite coordination must be applied. For global coverage, handover and management policies become more critical under too low elevation because of natural barriers. Handover policies and management are well analyzed under [7] and [8].

Analysis of random coverage time in mobile LEO satellite communications is also well treated by [9].

This paper discusses the single LEO satellite coverage aspects as an overture to the global coverage. Some characteristics of LEO satellites are given followed by coverage geometry. Finally the results of coverage simulation under different attitudes from 600km to 1200km at low elevation are presented.

II. LEO Satellites and Coverage

Microsatellites in Low Earth Orbits (LEO) have been in use for the past two decades, mainly dedicated for scientific purposes. LEO satellites have very wide scientific applications, from remote sensing of oceans, through analyses on Earth's climate changes, Earth's imagery with high resolution or astronomical purposes. These satellites provide opportunity for investigations for which alternative techniques are either difficult or impossible to be applied. Thus, it may be expected that such scientific missions will be further developed in the near future especially in fields where similar experiments by purely Earth-based means are impracticable. Ground stations have to be established in order to communicate with such satellites. Ground stations can communicate with LEO (Low Earth Orbiting) satellites only when the satellite is in their visibility region.

Satellites in these orbits have an orbital period of around (90-110) minutes. For satellites this is a short flyover period, which means that the antenna at the ground station must follow the satellite very fast with high pointing accuracy. The contact communication time between the satellite and the ground station takes (5-15) minutes 6-8 times during the day [6]. The Hubble Space Telescope, for example, operates at an altitude of about 610 km with an orbital period of 97 minutes [6]. Every satellite (especially, microsatellite when is dedicated for scientific purposes) carries special instruments that enable it to perform its mission [5] (for example, a satellite that studies the universe has a telescope, a satellite that helps forecast the weather carries cameras to track the movement of clouds).

On other hand from the communication perspective the goal of the future communication systems is to provide high quality broadband services with global coverage [10]. The satellite constellation is a convenient network solution for real time global coverage. The constellation is a system of low Earth orbit (LEO) identical satellites, launched in several orbital planes with the orbits having the same altitude. The

satellites move in a synchronized manner in trajectories relative to Earth. The application of low Earth orbit satellites organized in a *constellation* is an alternative to wireless telephone networks. Satellites in low orbits arranged in a constellation, work together by relaying information to each other and to the users on the ground. If satellites within a constellation are equipped with advanced on-board processing, they can communicate directly with each other by line of sight using inter-satellite links (ISL) [5].

To provide a global coverage to a diverse user population a number of LEO satellite networks have been proposed and implemented. The LEO satellite networks can support both the areas with terrestrial wireline and wireless networks that lack any network infrastructure [7]. Nowadays Several LEO constellations (Globalstar, Iridium, Ellipso) are active and operational [7], [8]. For complete coverage of the Earth's surface some overlapping between the adjacent satellites is necessary, to keep the continuity of real time services [9]. The global coverage can be considered as an interoperable complementary networking process of multiple satellites organized in constellation, each of them contributing with its individual coverage. This is achieved because LEO satellites move with respect to a fixed observer on Earth surface, and along with satellite movement also the coverage area changes its position continually creating a coverage belt as in Fig.1. Satellites under the same attitude under different inclination make different belts, enabling global coverage. Individual satellite coverage for few LEO satellites is presented in Fig. 2.

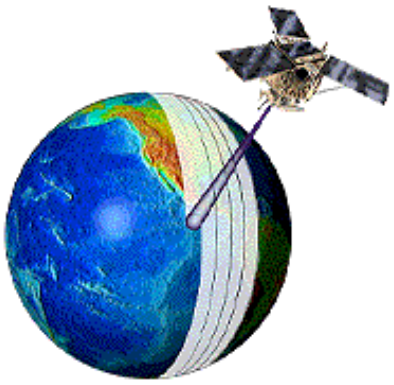


Fig. 1. Few LEO coverage area (Source: NOAA, 2009)

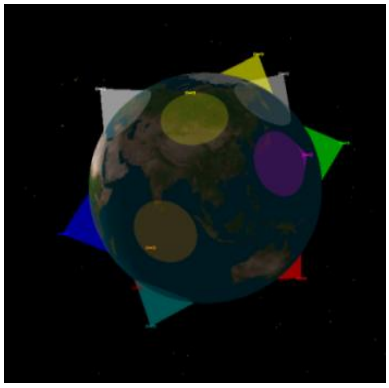


Fig. 2. Few LEOs coverage area (Source: NOAA, 2009)

When designing a satellite network some decisions such as the selection of the orbit parameters, coverage model, the network connectivity and routing model must be made. LEO satellite networks as Iridium, Teledesic, Globalstar, have architecture differences affecting the capabilities and services. The main problem for the global coverage is the handover process from one coverage area to another one. This is called satellite hand over process [7].

Different deterministic models for coverage time evaluation of low Earth orbiting satellite are developed. Models involve statistical coverage time assessments [8]. The analyses are particularly useful for probabilistic investigation of intersatellite handovers in LEO satellite networks [8]. The probability of service interruption and hand over mechanism becomes important for the overall system performance [9].

Achievements in antenna technology leaded to multibeam LEO systems where the footprint or coverage area is divided in many cells (multibeam arrays) in order to enhance frequency reuse policies. Applying space diversity policy is achieved frequency reuse inside a footprint. Handover from one cell to another is defined as cell handover. Particularly the interference problems have carefully to be treated [11] - [13].

III. LEO COVERAGE GEOMETRY

The position of the satellite within its orbit considered from the ground station point of view is defined by *Azimuth* (*Az*) and *Elevation* (ε_0) angles. The azimuth is the angle of the direction of the satellite, measured in the horizon plane from geographical north in clockwise direction. The range of azimuth is 0º to 360º. The elevation is the angle between a satellite and the observer's (ground station's) horizon plane. The range of elevation is 0º to 90º.

The coverage area of a single satellite is a circular area (Fig. 2) on the Earth surface in which the satellite can be seen under an elevation angle equal or greater than the minimum elevation angle determined by the link budget requirements of the system. The largest coverage area is achieved under elevation of 0º, but in order to avoid obstacles caused by natural barriers at too low elevation, usually for the link budget calculations it is determined the minimal elevation angle which ranges on (2-10)º. For simulation purposes of coverage it is considered the elevations up to 10 º.

The satellite's coverage area on the Earth depends on orbital parameters. Ground stations (GS) can communicate with LEO (Low Earth Orbiting) satellites only when the ground station is under coverage area (satellite footprint) as presented in Fig. 3.

The duration of the visibility and consequently the communication duration vary for each LEO satellite pass over the ground station, since LEO satellites move too fast over the Earth. Along with satellite, the footprint moves also, leaving the GS out of the footprint and consequently loosing the communication with the GS, as presented in Fig. 4.

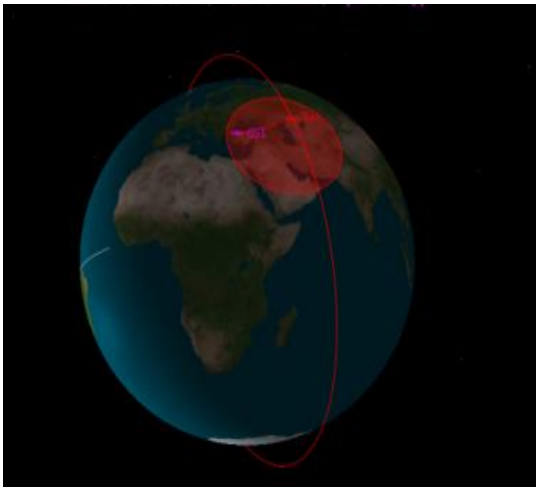


Fig. 3. The ground station (GS) under the LEO coverage area.

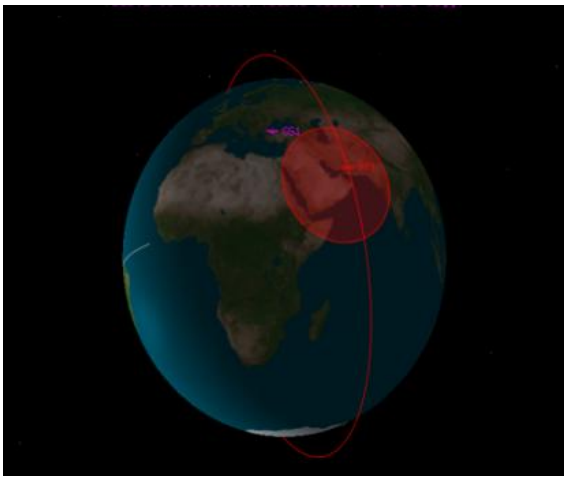


Fig. 4. The ground station (GS) out off the LEO coverage area.

The basic geometry between a satellite and ground station is depicted in Fig. 5. The points indicate the satellite (SAT), ground station (P), and then the third is the Earth's center. The line passing at point P represents horizon plane. The subsatellite point is indicated by T. Two sides of this triangle are usually known (the distance from the ground station to the Earth's center, $R_E = 6378x10^3$ m and the distance from the satellite to Earth's center-orbital radius). There are four variables in this triangle: ε_0 - is elevation angle, α_0 - is nadir angle, β_0 - is central angle and d is slant range. As soon as two quantities are known, the others can be found with the following equations [14]:

$$\varepsilon_0 + \alpha_0 + \beta_0 = 90 \quad (1)$$

$$d\cos\varepsilon_0 = r\sin\beta_0 \quad (2)$$

$$d\sin\alpha_0 = R_e\sin\beta_0 \quad (3)$$

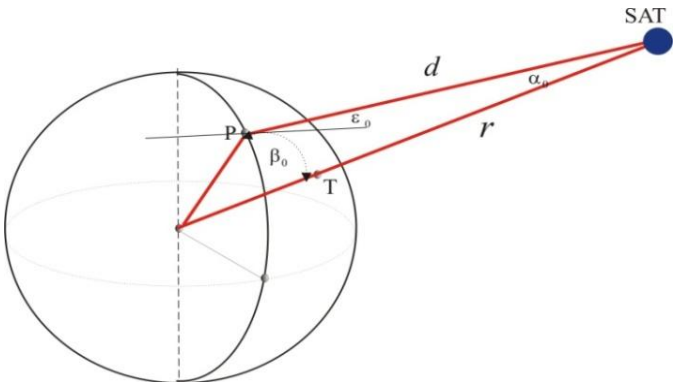


Fig. 5. Ground station geometry

The most needed parameter is the slant range d (distance from the ground station to the satellite). This parameter will be used during the link budget calculation, and it is expressed through elevation angle ε_0. Applying cosines law for triangle at Fig. 5 yields out:

$$r^2 = R_e^2 + d^2 - 2R_e d\cos(90 + \varepsilon_0) \quad (4)$$

Solving (4) by d, substituting, $r = H + R_e$ at (5) and applying (1), (2), (3) finally we will get the slant range as function of elevation angle ε_0 [14].

$$d(\varepsilon_0) = R_e\left[\sqrt{\left(\frac{H+R_e}{R_e}\right)^2 - \cos^2\varepsilon_0} - \sin\varepsilon_0\right] \quad (5)$$

H is the satellite attitude above the Earth's surface. Transforming Fig. 5 from the coverage point of view it looks like in Fig. 6.

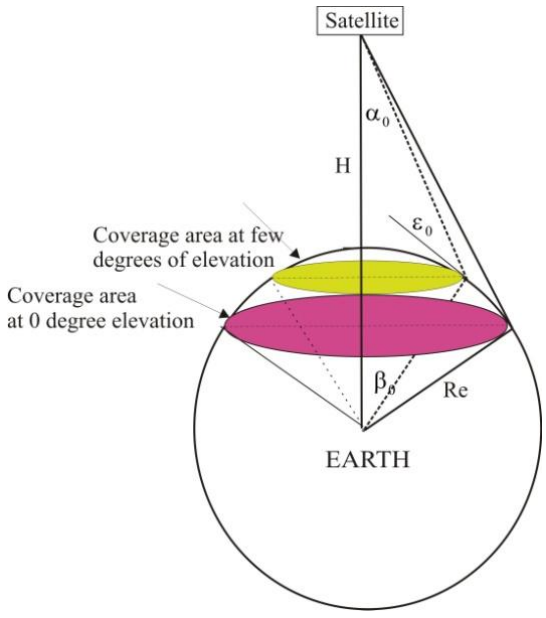


Fig. 6. Coverage geometry

In Fig. 6, there are two triangles. The larger one represents the case of full coverage under elevation of 0 °. Generally, from the smaller triangle applying sinus theorem yields out:

$$\frac{\sin\alpha_0}{R_e} = \frac{\sin(90+\varepsilon_0)}{R_e+H} \tag{6}$$

$$\sin\alpha_0 = \frac{R_e}{R_e+H}\cos\varepsilon_0 \tag{7}$$

The maximal coverage is achieved for $\varepsilon_0 = 0$, thus for known attitude *H*, easy is calculated the coverage angle for maximal coverage, as:

$$(\sin\alpha_0)_{MAX,H} = \frac{R_e}{R_e+H} \tag{8}$$

Similarly, for different elevations (ε_0) can be calculated α_0 and then based on (1) also β_0.

The surface of the coverage area depends on β_0 angle and it is [15]:

$$S_{Coverage} = 2\pi R_e^2(1-\cos\beta_0) \tag{9}$$

Usually, the satellite coverage area or the satellite's footprint is expressed (in percentage) as a fraction of the Earth's area.

$$Coverage(\%) = \frac{S_{Coverage}}{S_{Earth}} = \frac{2\pi R_e^2(1-\cos\beta_0)}{4\pi R_e^2} \tag{10}$$

$$Coverage(\%) = \frac{1}{2}(1-\cos\beta_0) \tag{11}$$

IV. LEO Coverage Simulation

Based on (7), (8) and (11) it is obvious that the satellite coverage strongly depends on elevation angle. To conclude about the coverage variation for low orbiting satellites at low elevation, the simulation for attitudes from 600km up to 1200km is further discussed.

For a given satellite attitude *H* and a given elevation angle ε_0 firstly should be calculated α_0, β_0 and finally the coverage based on (12). For attitudes of *H*=600km, 800km, 1000km and 1200km which are typical low orbit attitudes it is simulated and calculated the coverage area for elevation of (0-10)° by steps of 2°, and results are presented in Table I and Fig. 7.

Table I and Fig. 7 confirm the decrease of coverage area as elevation angle increases for the already defined attitude *H*, and the increase of the coverage area as attitude *H* increases keeping the fixed elevation.

TABLE I. COVERAGE ARE AS A FRACTION OF EARTH AREA.

Orbital Attitude [km]	**H 600 [km]**	**H 800 [km]**	**H 1000 [km]**	**H 1200 [km]**
Elevation (ε_0)	*Coverage* [%]	*Coverage* [%]	*Coverage* [%]	*Coverage* [%]
0°	4.30	5.60	6.80	7.95
2°	3.63	4.84	5.95	7.08
4°	3.05	4.16	5.21	6.22
6°	2.53	3.49	4.54	5.48
8°	2.08	3.01	3.91	4.75
10°	1.69	2.54	3.38	4.20

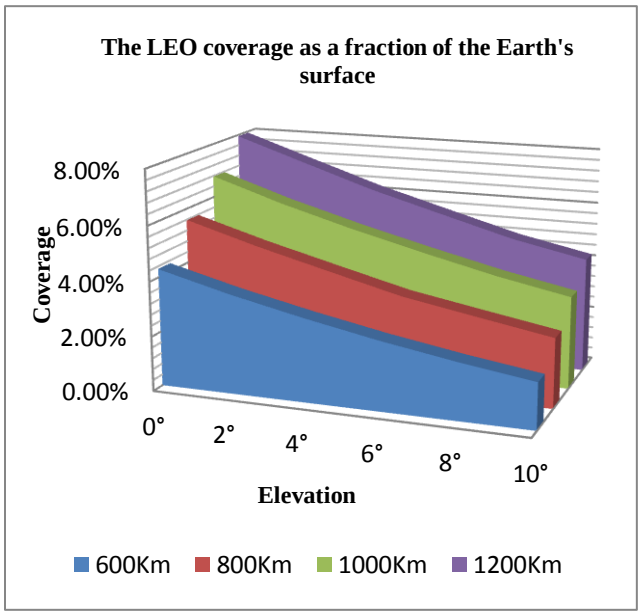


Fig. 7. Coverage area variation for different attitudes at low elevation

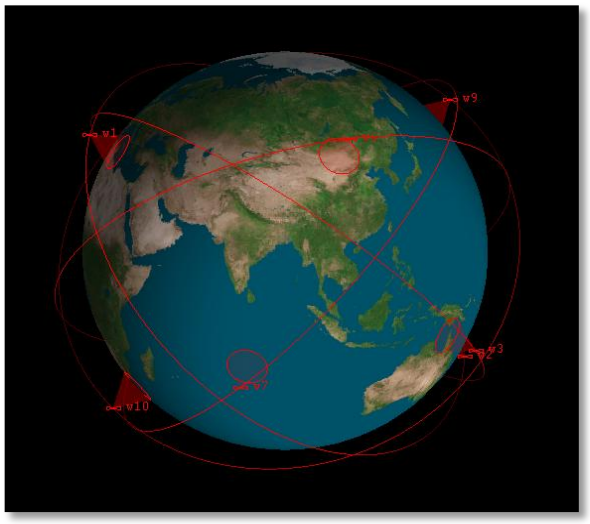


Fig. 8. LEO coverage area

Finally, in Fig. 8, applying satellite orbit analysis software is presented the case of simulated coverage area for synchronized orbits at an attitude of 600km for different inclination (few orbits) at elevation of 10 º, as the smallest coverage area stemmed from the simulation considered in this paper. Small circles in Fig. 8 represent LEO coverage area on the Earth's surface.

For the real time services the main problem due to global coverage remains the handover process from one coverage area to another one. For real time services it is too important the reliable communication what can be disturbed under to low elevation because of natural barriers. Our future work will be oriented on correlation of single coverage area with GIS (Geographic Information System) in order to have the exact information under which elevation the safe communication could be provided. The simulation tools will be essential part of foreseen future work.

CONCLUSION

The satellite's coverage area on the Earth depends on orbital parameters and usually is expressed (in percentage) as a fraction of the Earth's area. LEO satellites organized in constellations act as a convenient solution for real time global coverage. Satellite coverage strongly depends on elevation angle. The largest coverage area is achieved under elevation of 0º, but in order to avoid obstacles caused by natural barriers at too low elevation, usually for the link budget calculations it is determined the minimal elevation angle which ranges on (2-10)º.

Through simulation for typical LEO attitudes on range of (600-1200) km at low elevation of (0-10) º, it is confirmed that the fraction of Earth covered by satellites at appropriate attitudes is from 1.69% to 7.95%.

REFERENCES

[1] M. Richharia, (1999), "Satellite communication systems", McGraw Hill, New York, 1999.

[2] D. Roddy, "Satellite communications", McGraw Hill, New York, 2006.

[3] G. Maral and M. Bousquet, "Satellite communication systems", John Willey & Sons, Ltd, Chichester, England, 2002.

[4] www.answers.com/topic/van-allen-radiation -belt

[5] R. E. Zee, et al, "The MOST Microsatellite: A low cost enabling technology for future space science and technology missions", Canadian Aeronautics and Space Journal, 48(1), Canada, pg. 1-11, 2002.

[6] J.E. Oberright, "Satellite artificial", World Book Online Reference Center, World Book, Inc, 2004.

[7] I. F. Alyildiz, H. Uzunalioglu, M. D. Bender,Turkiye, " Handover management in Low Earth Orbit (LEO) satellite networks" Mobile Networks and Applications 4 (1999) 301 -310.

[8] P. Papapetrou, S. Karapantazis, F.N. Pavlidou, " Handover Policies in LEO Systems with Satellite Diversity", International Conference on Advanced Satellite Mobile Systems (ASMS 2003), 10-11 July, 2003, Frascati, Italy.

[9] Y. Seyedi, S. M. Safavi, "On the Analysis of Random Coverage Time in Mobile LEO Satellite Communications", Communications Letters, IEEE, Volume 16, Issue 5, MAY 2012.

[10] A. Botta, A. Pescape, "New generation satellite broadband Internet service: should ADSL and 3G worry", TMA 2013, co-lacted with IEEE INFOCOM 2013, April 2013, Turin, Italy.

[11] S. Cakaj, "Modulation Index Application for Satellite Adjacent Downlink Interference Identification", The 6th European Conference on Antennas and Propagation EUCAP 2012, IEEE, March 26-30, 2012 – Prague, Czech Republic, pp. 2000-2004.

[12] S. Cakaj, W. Keim, K. Malaric, "Intermodulation by Uplink Signal at Low Earth Orbiting Satellite Ground Station", 18th International Conference on Applied Electromagnetics and Communications, ICECom, IEEE, 12-14 October 2005, Dubrovnik, Croatia, pp. 193 - 196.

[13] S. Cakaj, K. Malaric, A. L. Schotlz, "Modelling of Interference Caused by Uplink Signal for Low Earth Orbiting Satellite Ground Stations", 17th IASTED International Conference on Applied Simulation and Modelling, ASM 2008, June, 23 –25, 2008, Corfu, Greece, pp. 187-191.

[14] G.D. Gordon, W.L. Morgan, 'Principles of communication satellites", John Wiley & sons, Inc. 1993.

[15] H. Curtis, "Orbital Mechanics for Engineering Students", Elsevier aerospace engineering series, pg. 55, 1998.

AUTHOR PROFILE

Shkelzen Cakaj has received his BSc and MSc degrees from Prishtina University in Kosovo. Since 2003 is cooperating with Institute for Communication and Radio – Frequency Engineering at the Technical University in Vienna, where he has prepared his Master Thesis related to the performance of the ground satellite station in July, 2004. He was awarded a PhD in area of satellite communication from Zagreb University in January 2008 with whom he has continued technical associations. He has attended courses on satellite communication and spectrum management at USTTI. He was awarded as Fulbright scholar researcher in 2009 at NOAA (National Oceanic and Atmospheric Administration) at Maryland, USA. He is the author of 45 papers published in worldwide conferences and journals; mostly IEEE. His area of interest is the performance of satellite ground stations for scientific satellites. He is working at Post and Telecommunication of Kosovo and lecturing satellite communications for master students at Prishtina University, Kosovo and Polytechnic University of Tirana, Albania.

Estimating Null Values in Database Using CBR and Supervised Learning Classification

Khaled Nasser ElSayed
Computer Science Department, Umm Al-Qura University

***Abstract*—Database and database systems have been used widely in almost, all life activities. Sometimes missed data items are discovered as missed or null values in the database tables. The presented paper proposes a design for a supervised learning system to estimate missed values found in the university database. The values of estimated data items or data it items used in estimation are numeric and not computed. The system performs data classification based on Case-Based Reasoning (CBR) to estimate loosed marks of students. A data set is used in training the system under the supervision of an expert. After training the system to classify and estimate null values under expert supervision, it starts classification and estimation of null data by itself.**

Keywords—DataBase(DB);Data mining; Case-Based Reasoning (CBR); Classification;Null Values; Supervised Learning

I. INTRODUCTION

Database is a collection of related data, to represent some aspects of the real world, sometimes called the mini-world or the universe of discourse. It has become an essential component of everyday life in modern society. In the course of a day, most of us encounter several activities that involve some interaction with a database.

RDBSs are the mostly database systems used today. These system organize databases in many relations. Each relation has data about certain entity type or class and consists of rows. Each row represent a record of entity or object. The state of the whole database will correspond to the states of all its relations at a particular point of a time.

Data Mining is an essential process where intelligent methods are applied in order to extract data patterns. Data mining algorithms look for patterns in data. While most existing Data Mining approaches look for patterns in a single data table, relational Data Mining (RDM) approaches look for patterns that involve multiple tables (relations) from a relational database [1].

In recent years, the most common types of patterns and approaches considered in Data Mining have been extended to the relational case and RDM now encompasses relational association rule discovery and relational decision tree induction, among others. RDM approaches have been successfully applied to a number of problems in a variety of areas, most notably in the area of bioinformatics. This chapter provides a brief introduction to RDM [2].

Knowledge discovery in databases (KDD), also called data mining, has recently received wide attention from practitioners and researchers. There are several attractive application areas for KDD, and it seems that techniques from machine learning, statistics, and databases can be profitably combined to obtain useful methods and systems for KDD [3].

The KDD area should be largely guided by (successful) applications. Theoretical work in the area is needed. A KDD process in which the analyzer first produces lots of potentially interesting rules, subgroup descriptions, patterns, etc., and then interactively selects the truly interesting ones from these [4]

The presented system uses CBR classification in estimation null values in DB. The basic idea is locating a classified case (a student object) in the system Knowledge Base (KB) as the most close case to the student case row which has a null value. After that, the system could estimate that null value using three methods and their average.

The weight of each attribute is varied, to represent its effect in the total mark. The total mark at any moment is a resultant of the already registered marks in the fields of the table. At any time, the weight of each attribute it is computed as output of dividing the attribute value for a student by the resultant of maximum values of all registered attributes for that course.

Section II present some survey on related work. While, section III, outlines the structure of database record used by the system. Section IV, explores the system knowledge base. Section VI explains training the system and system classification experiment and results, while section VII discuss the conclusion and future work.

II. RELATED WORK

A lot of research effort have been done in estimating null values in DB. The pioneers, Chen, in this area used a new method to estimate null values in relational database in [5]. They improved their method by creating fuzzy rule base in [6] and used genetic algorithms for generated weighted fuzzy rules in [7]. Then, they applied the automatic clustering algorithm for clustering the tuples in the relational database in [8]. Then, they presented a new method for estimate null values in relational database systems having negative dependency relationships between attributes in [9], where the "Benz secondhand car database" is used for the experiment.

Wang, C.H. Cheng, and W.T. Chang [10] utilized stepwise regression to select the important attributes from the database and a partitional approach to build the datacategory. They apply the clustering method to cluster output data. Also, Chen and Hsaio [11] and Cheng and Lin [12] utilized clustering algorithms to cluster data, and calculate coefficient values

between different attributes by generating minimum average error.

Jain and Suryawanshi [13] proposed an efficient approach for handling null values in web log. They used Tabu search–KNN classifier perform featureselection of K-NN rules.Also, C.H Cheng, J.R. Chang, and L.Y. Wei in [14] used adaptive learning techniques, based on clustering, toresolve the issue of null values in relational database systems. This study uses clustering algorithms to group data and calculates the degree of influence between independentattributes (variables) and the dependent attribute through an adaptive learning method.

Lee and Wang in [15] proposed a modular method for trying to processhigh-reliability relational database estimation, and thestructure of the proposed method can be composed of threephases, comprising partition determination, automatic fuzzysystem generation, and relational database estimation. While, Mridha and Banik used Noble evolutionary algorithm to generating weighted fuzzy rules to estimate null values [16].

Sadiq, S.A. Chawishly, and N.J. Sulaka in [17] proposed a hybrid approach for solving null values problem, it hybridize rough set theory with ID3 (Iterative Dichotomiser 3) decision tree induction algorithm. The proposed approach is a supervised learning model. Large set of complete data called learning data is used to find the decision rule sets that then have been used in solving the incomplete data. Then, the intelligent swarm algorithm is used for feature selection which represents bees algorithm as heuristic search algorithm combined with rough set theory as evaluation function[18].

III. Database Application

The proposed approach is tested in relational data base (DB) of university students. This database consists of many relations. Each relation is concerned of certain records of entity set. The target table is the STUDY table, shown in table 1, which concern of the remarks and grades of students in the registered courses.

Sometimes there missed or null values in a column of certain records in databases. As Example some remarks data of some student exams are missed. These null values might result from missing some exam grades or from non-entering mistakes.

As example, the estimation of null values is applied for a course has the assessments: two quiz (q1 and q2), five home works, a project, midterm exam, final exam. But it is possible to add or remove some assessment(s) to/from the proposed list of assessment(s). The STUDY table has those attributes, as shown in Table 1, which shows some records of student remarks.

The experiment is applied over marks data of the course, “Compilers Construction" in Computer Science department. Table 2, presents the universe of discourse of the attributes Home works, quizzes, MidTerm, Project, and Final Exam.

The attributes (column) of any database entry (SQL table) that have null values, are classified into four types, according to the reason and type of missing values or the ability of estimating the null values.

1) ***Type$_1$*** *is NullColumn,* where any column, like MedTerm as example, may have all of its values are null. This means that the column values are not inserted or computed yet.

2) ***Type$_2$*** *is NotEstimated*, where the attribute null values can't be estimated by any system. As example, the attributes: Student_num(St#), Name, Address, or Cours_Code.

3) ***Type$_3$*** *is Derived or Computed*, like Total. The attribute null value can be computed or imaged from another attribute(s).

The action in the first three types is running the program that computes or acquires those null data, or fill them by user.

4) ***Type$_4$*** *is Can-Estimated*, where a value of an attribute in certain row(student record) is missed. This null value can be estimated by the system.

TABLE I. Study Table With Actual Values of Homeworks, Quizzes, Midterm, Project, and Final Exams.

St#	Q1 /5	Q2 /5	H1/ 2	H2 /2	H3 /2	MTerm /20	H4 /2	H5 /2	Project /20	Final /40	Total /100
1	3	4	2	1	1	18	1	1	15	36	79
2	2	3	2	2	2	18	2	1	15	38	83
3	1.5	2.5	2	2	2	1	2	1	15	39	66.5
4	5	4	2	1	1	11	1	1	15	2	38
5	5	5	2	1	2	17	1	2	17	24	71
6	3	2	2	1	2	17	2	1	17	26	70
7	4	3	0	2	1	18	2	1	17	2	46
8	2	3	0	2	2	18	1	2	17	28	73
9	1	1	0	2	2	18	2	2	12	0	39
10	3	3	1	1	1	19	2	2	12	0	41
11	5	4	1	2	2	19	1	2	12	35	78
12	3	5	2	1	1	18	1	1	12	37	78
13	4	5	2	2	1	7	1	2	0	8	28
14	4	5	2	1	2	17	2	2	0	36	67
15	4	3	2	2	1	16	2	1	0	33	60
16	4	3	2	1	2	16	2	2	0	32	60

TABLE II. Universe of Discourse For Home Works, Quizzes, Midterm, Project, and Final Exams

Attribute Name (Assessment)	Minimum Value	Maximum Value
Home Works (H_1,H_2,H_3,H_4,H_5)	0	2
Quizzes (Q_1,Q_2)	0	5
MidTerm Exam	0	20
Project	0	20
Final Exam	0	40

The proposed method estimates null values in all column of type 4, based on the values of the known marks in the database. Thus, the known and estimated values are numeric values. Then system computes the total remark.

IV. Knowledge Representation

The system should acquires basic knowledge needed to build its KB, shown in Fig. 1. This process trains the system to

learn classification and estimation under the supervision of the expert. Fig. 2 demonstrates the algorithm for this process.

Each student object is scanned to be classified is stored as a case. Each case is described by its attributes of certain row in Table 1. Values of these attributes will be used in classification (category) of student objects. The category gives impression about the level of student objects related to it. It refers to the range within which their total resultant of registered attributes grades divided by the total of maximum marks of those attributes, in certain course. It has actual categories like: APLUS, A, BPLUS,B,……..,FAIL, LOWFAIL.

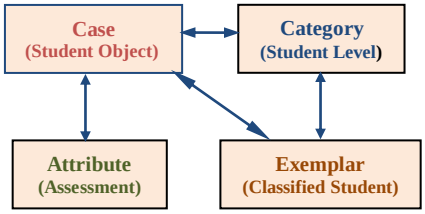


Fig. 1. System Knowledge Base

Each classified student object is related to a category. It is known as an exemplar of that category. It is represented as a combination of values of assessments attributes. There is no restriction on number or names of categories and exemplars.

Student object, Student Level, Attribute, and exemplar are represented as C++ classes. These classes and their relationships represent the knowledge base of the proposed system.

V. Training and Supervised Learning Process

At running the system for the first time, it reads the student objects (rows of Table) and checks there attributes for null values. Then, it gives report of null value types according to the preliminary classification given in section III. Also, it specifies the rows which has null values to be estimated later as described in section VI.

For each attribute a scale of possible values is determined. The combination of all possible attribute values defines all possible marks states within this description. The task is to classify each student object's state.

When a student objects (cases) are scanned by the system - for the fist time - to classify, it can do nothing. It has no categories and no classified exemplars to match. It'll ask for help from the expert to classify and clarify reason for that classification. First distinct cases will be classified by the expert and added to KB as exemplars. Those exemplar are related to new created categories.

At reading a new row of student object (unclassified case), the system will start classification process to specify a category from KB categories, based on values of its attributes. If the category is not in the KB yet, it will ask the expert to create new one, and name it. Categories names are listed section IV. Within each category, there will be many exemplar, each has its level. This level should be within the space of the category.

Exemplar level =*sum of actual values of all encountered attributes*/*sum of maximum values of all encountered attributes*.

For a new case, the system looks up for a similar exemplar to it. If it finds a category, it consult its suggestions to the expert. If the expert accepts, the new case is related to the category and a new exemplar is created if expert want. While, if the expert refuses that classification, or the system fails to find a category, it asks the expert to explain why? And classify himself and give reasons. Then a new category and an exemplar (new case) related to that category are created.

The expert may classify the new case to an existing category, or even a new one. The Algorithm of system training and classification is shown in Fig. 2. Supervised learning will continue in the estimation process, as seen in the next section.

Discard column of $type_1$, $type_2$, $type_3$, and $type_4$.
For each row of complete registered data Do

1. **Read actual and maximum values for each attribute of encountered column.**
2. **Compute the Total of Actual Values (TAV).**
3. **Compute the Total of Maximum Values (TMV).**
4. **Compute Case Level (CL) = TAV/MTV.**
5. **If there is no Category, its scope include CL, Consult expert to get name and scope of new Category, and Create it.**
6. **Exemplar level (EL) = CL.**
7. **Add new Exemplar related to the specified category.**
8. **Ask Expert for Classification reason and Learn.**

Fig. 2. Algorithm of System Training and Classification.

VI. Estimating Null Values

A. System Classification of Student Object

Mainly, the use of CBR classification is for locating the most close case (exemplar) to the student case which has null data. Student object of null value is the object has to be classified and assigned to a certain class (category). It is constructed from its marks of assessments (attributes) and their weight. It is clear that an expert Instructor uses that knowledge to characterize a marks condition. Assuming that in order to make preliminary conclusions the expert uses a finite number of marks of assessments.

Each attribute has a weight, based on the its space of minimum and maximum marl value that can be assigned to it, compared with the total of values all registered attributes for students.

Some attribute may be not available at certain moment for a course. There may an assessment canceled, or not held yet. So, the N/A attributes should excluded from the list of attributes that describe a student objects (cases) for a while, until be included in the DB. This happens as done with column of $type_1$, $type_2$, and $type_3$.

When a new case (student object), with null value in one of its attributes, is found out, the system start its classification process. It looks up for a category for that case and discovers the most similar exemplar (classification) to that case. If it fails, it asks for expert classification. While, if it successes, it starts estimating of the null value for the current classified student (case). The Algorithm of system classification and estimation is presented in Fig. 3.

After classifying the student object to be related to certain category, the system retrieves the exemplars related to the same category. It might use one of four methods to estimate null values.

1. **Read actual and maximum values for the student row where it has a null value.**
2. **Compute the Total of Actual Values (TAV).**
3. **Compute the Total of Maximum Values (TMV).**
4. **Compute *Case Level (CL) = TAV/MTV*.**
5. **If there is no Category, its scope include CL, Consult expert to get name and scope of new Category, and Create it.**
   Else locate category and all exemplars related to it
6. **Estimate Value for null Value using three estimation methods (Est_1, Est_2, and Est_3).**
7. **Compute the *EstAvg= (Est_1, Est_2, and Est_3)/3*.**
8. **Consult Estimation Values (Est_1, Est_2, Est_3, and EstAvg) to the Expert, to select one.**
9. **Ask Expert for selection reason, and Learn**.

Fig. 3. Algorithm for System Training and Classification.

Value of null attribute A in the current student record is estimated as any of the following methods:

1) The opposite value of the same attribute A in the most similar exemplar.

2) The average of all opposite values for attribute A in all exemplars related to the classification category.

*3) The average of level of all exemplars related to the classification category * maximum value of that attribute (out of marks).*

4) The average of the results from 1,2 and 3.

Then, the system offers its estimated values to the expert, to get his selection and guidance. The expert should choose one of them or refuse all. For all chooses, the system ask the expert for reasons of his decision.

Most of times, the expert reason was that the selected method is suitable for the nature, weight, and difficulty of each assessment (attribute). Next times, the system will use this knowledge to choose the method itself. Comparing results of estimating for assumed null values attribute will explain next. Finally, the system calculates the average of all methods. As seen next.

B. Experiment Results

Assume that there are n records (R_1, R2,….,R_n) in the STUDY table of the database, where the value of the attribute "MidTerm" of the record R_i is "R_i.MidTerm", as example. Also, assume that the estimated values of Ri.Midterm are ER_i.MidTerm (method1, method2, method3, method4). To estimate the value of the missed MidTerm value, those four values are estimated according to the four methods listed above.

Referring to the table STUDY showed in table 1, and assuming that there is null value in a certain records. five assumptions will be tested, while Table 3, collects results of the following experiment to estimate null values in an attribute of five columns of different records:

1) The record of 15th student record has null value in the column of MidTerm, while other attributes are given their values.

2) The record of 5th student record has null value in the column of homework H1, while other attributes are given their values.

3) The record of 8th student record has null value in the column of Final Exam, while other attributes are given their values.

4) The record of 10th student record has null value in the column of quiz Q1, while other attributes are given their values.

5) The record of 12th student record has null value in the column of Project, while other attributes are given their values.

TABLE III. RESULTS OF THE EXPERIMENTS

Experiment	Method1 value	Method2 Value	Method3 value	Average Value	Actual Value
ER_{15}.MedTerm	16	16	12	14.66	16
ER_5.H1	2	1	1.41	1.47	2
ER_8.FinalExam	24	25	18.5	22.53	28
ER_{10}.Q1	4	3.25	1.88	3.04	3
ER_{12}.Project	12	13.5	13	12.83	12

As seen in table 3, there is no method is preferred to applied for all attributes. While, the average of all estimation process, is somehow reasonable and applicable. Also, it is noticed that if the number of rows increases, the precision of the estimation will increase also.

VII. CONCLUSIONS

This paper presented a supervised learning system for estimating null values found in the database. The system performs data classification based on CBR-based classification to estimate missed marks of students. A moderate data set is used in training the system under the supervision of an expert, then the system start classification of objects that have null values using four methods. It is found that the average of the estimated values is more reasonable and applicable. In future, improvements will be applied to increase the precession of estimated values. Bigger training data set will be used in training the system to improve precision.

Also, the task of estimation will be enlarged to enable the system to estimate a multiple null values not only one null value in the in the same record.

REFERENCES

[1] J. Han & M. Kamber “Data Mining Concepts and Techniques”, 2nd edition, The Morgan Kaufmann Series in Data Management Systems Series Editor: Jim Gray, Microsoft Research Data Mining, ElServierInc, 2006.

[2] S. Džeroski “Data Mining and Knowledge Discovery Handbook”, Part 6, 887-911, Springer, 2010.

[3] H. Mannila& H. Toivonen “Levelwise Search and Borders of Theories in Knowledge Discovery”, Data Mining and Knowledge Discovery 1, 241–258, Kluwer Academic Publishers, Manufactured in The Netherlands, 1997.

[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, R. (Eds.), “Advances in Knowledge Discovery and Data Mining”, Menlo Park, CA: AAAI Press, 1996.

[5] S.M. Chen and H.H. Chen, "Estimating Null Values in the Distributed Relational Databases Environment", Cybern, Syst., Vol. 31, No. 8, pp. 851-871, 2000.

[6] S.M. Chen, S.H. Lee, and C.H. Lee, "A New Method for Generating Fuzzy Rule from Numerical Data for Handling Classification Problems", App. Art. Intell., Vol. 15, No. 7, pp. 645-664, 2001.

[7] S.M. Chen and C.M. Huang, "Generating Weighted Fuzzy Rules from Relational Database Systems for Estimating Null Values using Genetic Algorithms", IEEE Transactions on Fuzzy Systems, Vol. 11, No. 4, pp. 495-506, August 2003.

[8] S. M. Chen and C. M. Huang, “A new approach to generate weighted fuzzy rules using genetic algorithms for estimating null values,” Expert Systems with Applications, vol. 35, no. 3, pp. 905-917, October 2008.

[9] S.M Chen and S.T. Chang, "Estimating Null Values in Relational Database Systems Having Negative Dependency Relationships Between Attributes", Journal Cybernetics and Systems, , Vol. 40, No. 2, pp. 146-159, February 2009.

[10] J.W. Wang, C.H. Cheng, and W.T. Chang, "Partitional Approach for Estimating Null Value in Relational Database", Springer-Verlag AI 2005, LNAI 3809, pp. 1213-1216, 2005.

[11] S.M. Chen and H.R. Hsiao, "A new method to estimate null values in relational database systems based on automatic clustering techniques", Elsevier Inc., Information Sciences 169, pp. 47-69, 2005.

[12] C.H. Cheng and T.C. Lin, "Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value", ICICIC, 2007, Innovative Computing, Information and Control, International Conference on, Innovative Computing, Information and Control, International Conference on 2007, pp. 81-89.

[13] Y. K. Jain and V. Suryawanshi, "A New Approach for Handling Null values in Web Log Using KNN and Tabu Search KNN", International Journal of Data mining &Knowledge Management Process (IJDKP), Vol. 1, No. 5, pp.9-19, September 2011.

[14] C.H Cheng, J.R. Chang, and L.Y. Wei, "ADAPTIVE-CLUSTERING BASED METHOD TO ESTIMATE NULL VALUES IN RELATIONAL DATABASES", International Journal of Innovative Computing, Information and Control(ICIC), Vol. 7, No. 1, pp. 223-235, January 2011.

[15] S.J. Lee and H.S. Wang, "A Dynamic Modular Method for Estimating Null Values in Relational Database Systems", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), Vol. 1, pp. 249-257, 2009.

[16] M.F. Mridha and M. Banik, "Performances of Estimating Null Values using Noble Evolutionary Algorithm (NEAs) by generating Weighted Fuzzy Rules", International Journal of Computer Applications, Vol. 11, No. 9, pp. 30-35, December 2010.

[17] A.T. Sadiq, S.A. Chawishly, and N.J. Sulaka, "Intelligent Methods to Solve Null Values Problem in Databases ", Journal of Advanced Computer Science and Technology Research, Vol. 2, No. 2, pp. 91-103, June 2012.

[18] A.T. Sadiq, M.G. Duaimi, and S. A. Shaker, "Data Missing Solution Using Rough Set Theory and Swarm Intelligence", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3, PP. 1-16,2013.

AUTHOR PROFILE

The Author is Dr. Eng. Khaled N. ElSayed. He was born in Cairo, Egypt 9 Oct. 1963. He have got his PhD of computers and systems fromFaculty of Engineering, Ain Shams University, Cairo, Egypt, 1996.

He has worked as an associate professor of computer science, in Umm-AlQura Uni. in Makkah, Saudi Arabia since 2006. Artificial Intelligence is his major. His interest research is Distant Education, E-Learning, and Agent.

Dr. Khaled N. ElSayed translated the 4th edition of “Fundamentals of Database Systems”, RamezElmasei and Shamkant B. Navathe, Addison Wesley, fourth edition, 2004, published by King Saud University, Riyadh, Saudi Arabia, 2009. He is also the author several books in programming in C & C++, Data structures in C& C++, Computer and Society, Database Design and Artificial Intelligence.

An Experience of Taiwan Policy Development To Accelerate Cloud Migration

Sheng-Chi Chen
Department of Management Information Systems,
National Chengchi University
Taipei, Taiwan

***Abstract*—Developing cloud computing is a key policy for government, while convenient service is an important issue for people living. In the beginning of 2010, the Taiwan Government has launched a "Cloud Computing Development Project", and has devoted to service planning and investment activities. At the end of 2012, in a three-year comprehensive review and suggestion adoption from public and private sectors, the Taiwan Government adjusted the policy and rename as "Cloud Computing Application and Development Project".**

From the perspectives of government application, industry development, and cloud open platform, this study describes how the vision drive goals and thinking push forward strategies. In the process of government and industry collaboration, it is progressively created value for cloud services. The Cloud Computing Project Management Office acts a key role as policy advisor, matching platform, and technical supporting to the achievements of (1) policy assessment and strategy enhancement; (2) construction of cloud open platform to the demand and supply linkage; (3) innovation and integration planning for government service application, leading to industry development.

Keywords—Cloud Computing; Action Research; Project Management Office

I. INTRODUCTION

The development of the cloud computing industry re-shapes the global IT industry. Cloud computing technology and its service applications are taking off around the world. To seize the initiative and ensure future competitive advantages in the cloud computing market, Taiwan is adapting its currently-established hardware manufacturing foundation and extending it to its IT industry. The Cloud Computing Development Project, initiated by the Taiwan government in early 2009, includes 15 cloud computing projects. After three years of effort, by the end of 2012, the proposals were revised to accommodate the diversity of domestic demand and technological competition from international industries. The cloud applications by the government shall be transparent to the general public and shall lead the way for the cloud computing industry's development in Taiwan. The cloud platform, named Cloud Open Lab, is offered as a supply-demand channel in between governmental agencies and the hardware/software vendors in cloud computing applications. Generally speaking, additional governmental procurement projects for hardware equipment are not recommended. Furthermore, an evaluation policy for the government cloud computing project is provided, for "Value to citizens" and "Economy to the industry" as the planning and implementation target guidelines. At the same time, the Cloud Computing Development Project has been revised to become the Cloud Computing Application and Development Project. This research discusses how the two cloud computing projects successively lead to accelerate cloud migration of the ICT development in Taiwan.

II. RELATED WORKS

A. Cloud Computing

Gartner pointed out that cloud computing is a computational mode that is large-scale and provides IT capabilities, to be accessed by multiple external users over the Internet [2][5]. The National Institute of Standards and Technology (NIST) defines: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction." The NIST definition lists five essential characteristics of cloud computing: on-demand self-service, broad network access, resource pooling, rapid elasticity or expansion, and measured service. It also lists four "deployment models" (private, community, public and hybrid) that together categorize ways to deliver cloud services, and three "service models": Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [4][8].

B. Government Policy

Cloud computing has been regarded as one of the crucial industries in Taiwan. With cloud computing having gradually emerged as the mainstream of future ICT applications, the governments of advanced economies throughout the world have been working actively to formulate strategies for cloud computing development. Government agencies will be able to benefit from cloud computing to improve efficiency and reduce costs, further upgrading and restructuring the industry services with core competitiveness to drive market growth in both domestic demand and export sales. The U.S., which already possesses a substantial cloud computing industry, has been focusing on enhancing government efficiency and reducing costs; the European Union (EU) and South Korea have made the development of overseas export markets the main focus of their cloud computing strategies, while Japan has positioned cloud computing as a key tool for strengthening the competitiveness of the Japanese ICT sector. Common features seen in many of the cloud computing development policies

formulated by these various governments include the use of cloud computing as a foundation for developing mobile services that provide an improved user experience, creating value through open access to data, formulating relevant standards and regulations, simplifying government procurement procedures, and utilizing big data analysis techniques to help improve government performance and promote commercial development [1].

C. Industry Development

Google and Amazon.com are the two leaders in the development of global cloud computing applications. Google started out as a web search service and branched out to e-mail, online video, maps, social networking, and various online software services. Amazon.com's core business is in e-commerce, selling books over the Internet, and then expanded to multimedia, software, electronics, and household items. This created a new business model with indicative significance for the cloud computing services. Google and Amazon.com are based on the Internet and software services, starting out by providing software as a service (SaaS) and then scaling up their business with large-scale data centers to provide computing resources such as platform as a service (PaaS) and infrastructure as a service (IaaS). They are the leaders in cloud computing services. Major global players such as Microsoft, IBM, VMware, AT&T, and Apple have also developed their own cloud computing services to further drive innovation into the emerging cloud applications [1].

III. Research Method

Action research refers to research where the participant solves an immediate problem with active participation in the situation whilst conducting research for the solution, thereby improving the working efficiency [3] [7]. Action research allows the researcher to observe and describe the conditions in a practical work setting, wherein the researcher can participate in the change and continue to evaluate the process. First a variable is identified, a course of action is defined, and then the issue is monitored for continuous evaluation. Action research composes of a cycle of steps composed of planning, action, fact-finding, and reflection. The immediate response from the scenario can be verified after each action is carried out [6][9].

In the study, the researcher participated in the actual process of policy formulation and platform establishment, using interviews and observation to collect comprehensive data, and interacting with related parties, to examine the process of promoting cloud computing industry development and the analysis of relevant problems in government, employing actual participation and continuing improvement to explore these issues. As shown in Figure 1, this paper is an action research from the perspective of an aide of the Cloud Computing Project Management Office and investigates the implementation and challenges of the Cloud Computing Development Project and Cloud Computing Application and Development Project, wherein the discussion stems from actual participation and improvements made to the projects [1].

A. Cloud Computing Development Project

In 2009, the Taiwanese government positioned cloud computing as a key strategic industry the development of which needed to be prioritized. This was followed in 2010 by the launching of the Cloud Computing Development Project, which sought to use cloud computing to link together the IT hardware, software and information services segments within Taiwan's ICT sector through a total of 15 cloud computing development applications.

B. Cloud Computing Application and Development Project

In 2012, responding to changes in market demand and intensified international competition in terms of technology, etc., the overall strategic direction and objectives of the Cloud Computing Development Project were revised to emphasize value creation and production value, and the Project was renamed the Cloud Computing Application and Development Project, shown in Table 1. It was anticipated that, through the development of public-sector cloud computing applications, it would be possible to stimulate the continued development of Taiwan's cloud computing industry, while also planning the establishment of cloud computing service platforms to serve as matching mechanisms for supply and demand.

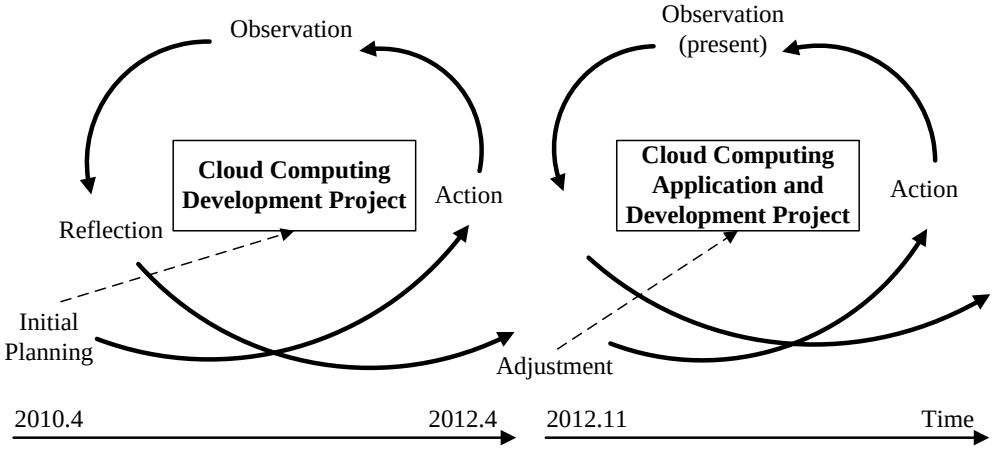


Fig. 1. Action Research Process

TABLE I. THE ANALYSIS OF POLICY EVOLUTION IN TAIWAN

	Cloud Computing Development Project	Cloud Computing Application and Development Project
Period	3.5 years (2009.4-2012.10)	1.5 years (2012.11-present)
Target	G2B, B2B	G2B, B2B, G2C
Strategy	● Develop advanced technology and infrastructure ● Popularize concepts and demonstrate new applications	● Promote the valuable applications for citizens ● Build the foundation for the system software ● Energy efficiency ● Develop the expertise for applications ● Leverage local infrastructure
Framework	● Governance: Committee and PMO established ● Supply: Leverage C4(cloud, connectivity, commerce, client) industry chain ● Demand: e-Government to new G-Cloud ideas	● Dual focus (application & industry) and bridging by PMO ● Supply: technology and promotion ● Demand: roll-out unqualified G-Cloud targets
IT artifact supported	N/A	Cloud Open Lab for matchmaking
Application Outcomes	● 3 specifics areas ● 10 applications	● 5 specifics areas ● 10 applications

With the developments taking place in cloud computing technology, the allocation of resources by both the government and the private sector to cloud computing technology development – including collaboration between Taiwanese corporations and research institutes to develop cloud computing application platforms, building on existing IT hardware to design and develop innovative new cloud computing applications – will drive the development of the cloud computing sector and the environment that supports it; at the same time, as more Taiwanese firms begin to develop overseas markets for cloud computing applications, these initiatives will contribute to the continued growth of the Taiwanese economy as a whole [10].

IV. A KEY ROLE：PMO

The main functions of Cloud Computing Project Management Office are to help domestic companies participate in government projects, link the cloud computing industry and government application services together, and carry out the inter-ministerial integration of services to businesses and the people as well as the use of government resources. The office provides technical and administrative leadership to cloud computing initiatives. The Cloud Computing Project Management Office proposed the revised project structure to address the international market demands and technical competition. The Cloud Computing Development Project has undergone a constant roll-out process after its introduction in 2010, with a series of revisions aimed at adjusting the project accordingly for continued innovation and matching government policy to industry needs.

Initially, the revision called for attention to the supply, demand, and administrative aspects of the previous projects to propose the two major adjustment policies. The first is to improve the supply and demand through 5 major policies: "Promote the valuable applications for citizens," "Build the foundation for the system software," "Energy efficiency," "Develop the expertise for applications," and "Leverage local infrastructure" to promote the relevant practices.

The second is to strengthen the governing policy and secure the mechanism for promoting supply and demand by adding the additional concept of supply-demand matching as provided by the Cloud Open Lab. The government shall lead the industry with official policies, to truly expand and develop the cloud industry in Taiwan. Based on the two major policies, the structure of the project is revised with the following three key focuses:

A. *Dual-focus triple-stage framework*

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

1) Develop a strategy: Develop the 5 major driving policies based on the project guidelines.

2) Plan an evaluation policy: Develop the evaluation policy based on the cloud computing guidelines.

3) Agency integration project: Integrate the agency development projects with key focus on government applications.

B. *Triple-stage process for establishing the supply-demand matching platform*

Implement the Cloud Open Lab supply-demand matching service:

1) Planning and implementation of the platform: Energize the IT vendors of Taiwan through the government cloud applications, to verify, validate, and certify the supply and demand scenario.

2) Collect industry solutions: Provide a channel for business opportunities and create cloud solutions developed locally in Taiwan.

3) Government promotion and explanations: One-stop service to reduce initial cost of the cloud application.

C. Leading with triple-stage government application

Government applications to lead industry development and demonstrate the performance to the public:

1) Cross-agency communication and coordination: Government agencies are invited to develop a solution and strategy, and reach a consensus for the project.

2) Assist agencies in the project proposal: Government agencies are invited and give counseling for project proposals.

3) Planning, evaluation, and integration: The proposed projects from all agencies are evaluated.

V. Result

With the two major adjustment directions and three key areas of focus, the various items of the Cloud Computing Application and Development Project can be revised individually. This leads to a new wave of development for cloud computing in government applications and creates an industry roadmap in Taiwan. The performance and value created by the project can be investigated from two aspects:

A. Dual-focus revised cloud computing project for multi-fold advantages

The Cloud Computing Application and Development Project approved by the Executive Yuan shall focus on the dual-focus framework of "Government applications" and "Industry development." The "technical value" and "social value" shall also highlight the innovative value of cloud computing, thus the driving force and strategy behind the industry development shall be government applications.

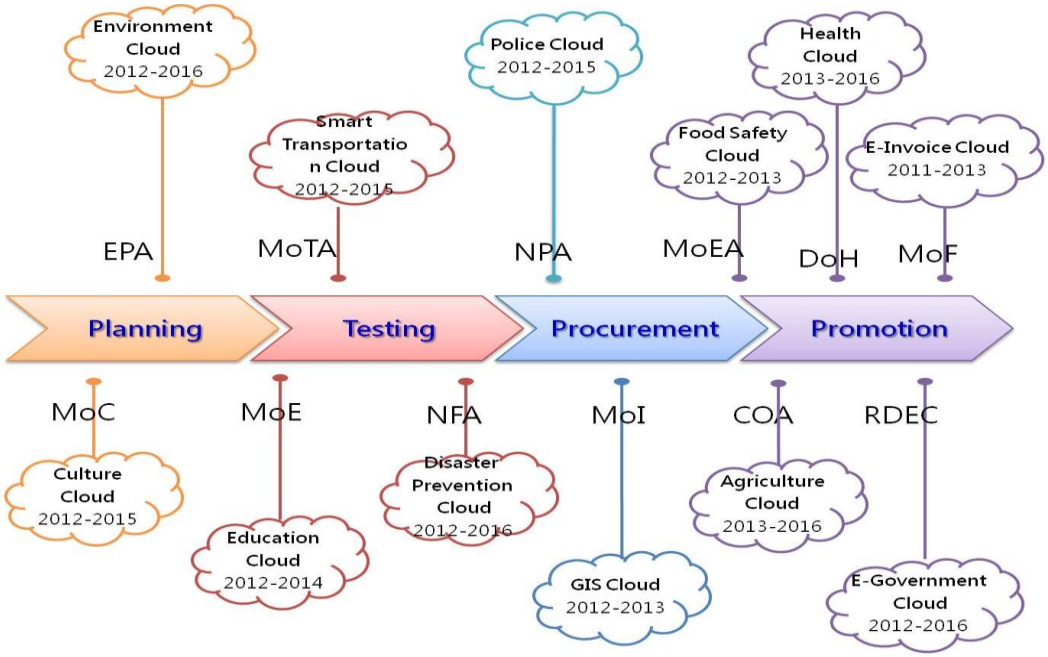


Fig. 2. Development Status of Taiwan Government Cloud

Government agencies shall be invited to discuss the 5 specifics areas of "Food and health safety," "Police and traffic," "Environmental resources and hazard prevention," "Education and culture," and "Public infrastructure" government applications. As shown in Figure 2 and Table 2, a target of 10 government applications shall be made available to the general public, with a target of 10 million people served by the government cloud services.

B. Cloud Open Lab for supply-demand matching

The Cloud Open Lab is a centralized hub for cloud solutions that provides supply-demand matching services, promotes industry R&D efforts, and links government applications to the industry services. IT venders of Taiwan are strengthened by the demand driven by the government cloud services. The Cloud Open Lab developed brings together a variety of different cloud-based solutions to reduce search costs for private-sector firms and implement open purchasing standards; it provides commercialization environment resources and testing opportunities so that firms can implement service feasibility testing in advance, thereby reducing unnecessary hardware procurement and development costs, and increasing the probability than any given development project will be a success. The main functions of the platform, shown in Figure 3, are as follows:

1) Proof-of-concept (POC): Recruiting private-sector firms' existing cloud computing resources (supply side) to target particular cloud-based applications (demand side), providing matching and preliminary concept feasibility verification testing service.

2) Verification: Providing verification testing services for cloud-based applications (demand side) that verify the special features of these applications; providing verification testing services for cloud-based service level agreements (supply side) that verify conformity with openness criteria.

3) Certification: Providing certification and testing services, involving formal, signed documentation and/or marks or logos, in accordance with relevant cloud computing standards and rules.

TABLE II. SUMMARY OF TAIWAN GOVERNMENT CLOUD

Application	Description
E-Invoice	Establishing E-Invoice systems, promoting the development of paperless invoice systems, and replacing conventional paper invoices.
Transportation	Taking and expanding information collection to provide a wider range of transportation information services and enhance transportation information service quality.
Police	Building "M-Police" and "smart" case investigation capabilities, so as to improve the level of service the police provide for the general public and enhance the public's quality of policing.
Health	Establishing a cloud-based electronic patient records database that can be used for inter-hospital examination of patient records, while moving individual health data into the cloud.
Environment	Integrating cross-agency environmental big data, to provide services relating to residential environmental information, environmental and ecological resources, and environmental monitoring imaging.
Education	Integrating and utilizing existing educational resources, including e-learning resources, to create a friendly, "smart," invigorating environment in which students, teachers and parents can share resources and engage in smooth, natural exchange and interaction.
Culture	Building integrated artistic and cultural activities information services, and adopting a philosophy of open access to information, that all citizens have ready access to culture-related information.
Disaster Prevention	Integrating different types of information to provide weather, disaster, traffic, evacuation and other information and early warnings to citizens in a timely manner, thereby enhancing the effectiveness of disaster prevention and response management.
GIS	Making available the standardized, reliable, usable, frequently updated GIS information that government agencies and related organizations need for major national infrastructure projects.
Food Safety	Integrating food traceability platforms and completing the inter-agency integration, using food traceability records and monitoring to promote transparency in the food supply chain.
Agriculture	Establish agricultural production traceability, help agricultural enterprises to implement IT-enabled management, and enhance the competitiveness of the agricultural sector as a whole.

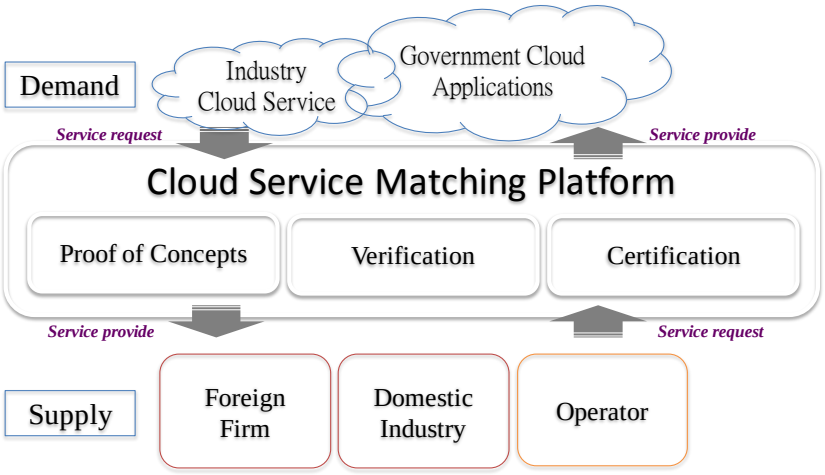


Fig. 3. Cloud Open Lab- A Connecting Bridge to Government and Enterprises

The Cloud Open Lab is a centralized resource for the industry that provides a training ground for driving R&D to discover new markets for the cloud computing industry of Taiwan. There are currently over 80 vendors in the Cloud Open Lab, with over 142 products on market shelves today. In terms of demand, the goal is to provide a one-stop service to reduce initial cost of the cloud application.

Government agencies shall be able to take advantage of the Cloud Open Lab to bridge the matching services and help test the government applications. Initially, a total of 9 agencies have joined the platform to trial run the government cloud applications on the vendor's products.

VI. CONCLUSION

With a first-class network environment, IT environment and IT hardware manufacturing sector, Taiwan is well-placed to develop cloud computing. The overall level of IT hardware manufacturing technology in Taiwan is extremely advanced, the broadband Internet access penetration rate is high, and a high percentage of Taiwanese business enterprises have achieved an impressive level of e-enablement.

Both the Cloud Computing Development Project and Cloud Computing Application and Development Project, the cloud computing strategy roadmap, outline the implementation of cloud computing in Taiwan to create a smart lifestyle and

kickstart the nation onto the path of becoming a technological powerhouse. The cloud industry has received strong official support with the government setting the pace for private investment and vendor to follow with the proprietary cloud computing infrastructure as the model platform. In accordance with the development strategy, the Cloud Computing Project Management Office is assisting vendors in participation of the government project to accelerate the development of the industry chain and launch e-government cloud services. The office, as a key role, is responsible for the planning and development of the cloud computing industry through comprehensive coordination, control, and execution. The office is tasked with assisting vendors in the investment and development of the government and corporate cloud services, thus boosting industry value and investment capacity in cloud computing applications.

VII. Further Research

Looking to the future, the office shall focus open data, government procurement and broadband infrastructure, to undertake planning for various cloud-based government services and encourage hardware manufacturers and software providers to from strategic alliances to pursue innovation, to promote the development of an ever wider range of cloud-based applications, and to re-engineer and streamline administrative procedures at all levels of government for better efficiency and transparency.

Most G-Clouds have moved to continuous improvement stage, some are still in the planning stage. Further researches can discuss obstacles and enablers of G-Cloud development, and the key factors to link the government and industry. From different G-Clouds and industry solutions development, researchers can offer various views to the cloud computing policy forming and implementation by the ethnographic research method.

Acknowledgment

The paper is, executed by Institute for Information Industry (III), supported by Ministry of Economic Affairs (MOEA), Taiwan. The author of this research, acting as a policy drafting staff, is involved in the progress of the cloud computing policy development in Taiwan (since 2010 - present).

References

[1] Chen S. C., “The Evolution of Taiwan Cloud Computing Policy: An Action Research,” The Proceeding of 2014 International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology (e-CASE & e-Tech 2014), pp. 483-489, April 2014.

[2] Daryl C. P., David M. S., Thomas J. B., David W. C., David J. C., Donna S., Rakesh K., Bruce R., “Five Refining Attributes of Public and Private Cloud Computing,” Gartner Research, 2009.

[3] Elliott, J., Action Research for Educational Change, Open University Press, Milton Keynes, 1991.

[4] Fang L., Jin T., Jian M., Robert B., John M., Lee B., Dawn L., “NIST-SP 500-292: NIST Cloud Computing Reference Architecture,” National Institute Standards and Technology, 2011.

[5] Frank E. G., Christopher M., Ellen D., Christina L., “How to Message Cloud Offerings and Not Get Lost In the Fog,” Forrester Research Inc, 2009.

[6] Kemmis, S., and Mctaggart, R., The Action Research Planner (3rd), Deakin University Press, Victoria, Australia, 1997.

[7] Lewin, K., “Action research and minority problems,” Journal of Social Issues, Vol. 2, No. 1, pp. 34-46, 1946.

[8] Michael H., Fang L., Annie S., Jin T., “NIST-SP 500-291: NIST Cloud Computing Standards Roadmap,” National Institute Standards and Technology, 2011.

[9] Wadsworth, Y., “What is Participatory Action Research?” Action Research International, Paper No.2. (website available at http://www.scu.edu.au/schools/gcm/ar/ari/p-ywadsworth98.html), 1998.

[10] Wu, C.C. and Chen S. C., “Migrating to the Cloud- A Review and Prospect of Taiwan ICT Vendors to Cloud Computing Market,” Proceedings of the 19th Conference on Information Management & Practice (IMP 2013), pp. 367-381, November 2013.

Educational Data Mining Model Using Rattle

Sadiq Hussain
System Administrator
Dibrugarh University
Dibrugarh Assam

G.C. Hazarika
Department of Mathematics
Dibrugarh University
Dibrugarh Assam

***Abstract*—Data Mining is the extraction of knowledge from the large databases. Data Mining had affected all the fields from combating terror attacks to the human genome databases. For different data analysis, R programming has a key role to play. Rattle, an effective GUI for R Programming is used extensively for generating reports based on several current trends models like random forest, support vector machine etc. It is otherwise hard to compare which model to choose for the data that needs to be mined. This paper proposes a method using Rattle for selection of Educational Data Mining Model.**

Keywords—Educational Data Mining; R Programming; Rattle; ROC Curve; Support Vector Machine; Random Forest

I. Introduction

Dibrugarh University, the easternmost University of India was set up in 1965 under the provisions of the Dibrugarh University Act, 1965 enacted by the Assam Legislative Assembly. It is a teaching-cum-affiliating University with limited residential facilities. The University is situated at Rajabheta at a distance of about five kilometers to the south of the premier town of Dibrugarh in the eastern part of Assam as well as India. Dibrugarh, a commercially and industrially advanced town in the entire northeastern region also enjoys a unique place in the fields of Art, Literature and Culture. The district of Dibrugarh is well known for its vast treasure of minerals (including oil and natural gas and coal), flora and fauna and largest concentration of tea plantations. The diverse tribes with their distinct dialects, customs, traditions and culture form a polychromatic ethnic mosaic, which becomes a paradise for the study of Anthropology and Sociology, besides art and culture. The Dibrugarh University Campus is well linked by roads, rails, air and waterways. The National Highway No.37 passes through the University Campus. The territorial jurisdiction of Dibrugarh University covers seven districts of Upper Assam, viz, Dibrugarh, Tinsukia, Sivasagar, Jorhat, Golaghat, Dhemaji and Lakhimpur. [1]

There are more than hundred numbers of Colleges/ Institutes offering TDC (Three Year Degree) Course affiliated/ permitted under the University. Since the number of students in the Arts Stream is larger in comparison to the other stream (B.Sc., B.Com., B.Tech. etc) we considered the data for the B.A. (Bachelor of Arts) course for our present study of educational data mining. The required digitized data are collected from Dibrugarh University Examination Branch for the affiliated colleges of the University B.A. programme from 2010 to 2013. This paper evaluates performance gender wise as well as caste wise of the students. The Colleges are categorized as Urban as well as Rural depending on their locations. In case of caste wise observations, the binomial operators are Urban and Rural.

There are several data mining tools and statistical models available. This paper focuses one which data mining tools shall be the best suited and what would be the statistical models for such knowledge discovery.

II. Literature Review

A. Data Mining

Data Mining detects the relevant patterns from databases / data warehouses using different programs and algorithms to look into current and historical data which can be analyzed to predict future trends [2]. It is very difficult for any organization to extract hidden patterns from the huge data marts and data ware houses without the help of data mining tools and programs. It is like searching for the pearls in the sea of data. This knowledge set is extremely useful in developing a knowledge support system and making important decisions regarding the future trends predictions.

Statisticians have used different manual techniques for the benefit of the business, predicting trends and results based on data over the years. The business houses had developed huge databases or data warehouses to become "data tombs". The data was never transformed into information. But with the help of data mining tools and algorithms now professionals from different areas may extract knowledge quickly and at ease.

B. Educational Data Mining

Data mining, often called knowledge discovery in database (KDD), is known for its powerful role in uncovering hidden information from large volumes of data [3]. Its advantages have landed its application in numerous fields including e-commerce, bioinformatics and lately, within the educational research which commonly known as Educational Data Mining (EDM) [4]. EDM is defined by The Educational Data Mining community website, www.educationaldatamining.org as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting, and using those methods to better understand students, and the settings which they learn in. EDM often stresses with the improvement of student models which denote the student's current knowledge, motivation and attitudes [5].

C. Rattle: A Data Mining GUI for R

The data miner draws heavily on methodologies, techniques and algorithms from statistics, machine learning, and computer science [6]. R programming language is a powerful tool for

data mining. Rattle (the R Analytical Tool To Learn Easily) provides GUI for the R programming environment. We have to use the library (rattle) and rattle () brings up the GUI for the programmers. Highly skilled Statisticians may efficiently use the R Programming Language. So, it is out of reach for many people without in depth knowledge of Statistics. But Rattle provides sophisticated GUI for data analysis and provides the necessary graphs with a click. Rattle provides another magnitude to the R programming and a platform for the novice data miners to work efficiently. Rattle's user interface provides an entry into the power of R as a data mining tool. [6]

D. ROC Curves Analysis

To determine a cutoff value, Receiver operating characteristic (ROC) curves is used in many areas. We may use the ROC curve for the selection of best suited models. In our educational data mining experiment, we use the ROC curve to determine the selection of model.

III. Experiments and Evaluation

A. The Data Set

We have included a small part of the Category and Gender based tables termed as Table 1 and Table 2 for which the suitable models needs to be selected. The Examination Branch of Dibrugarh University provides various College Codes for different Colleges under its jurisdiction. The field 'Appeared' means the number of candidates appeared for that examination and 'Passed' means the number of candidates passed for that particular examination. The field 'PassPercentage' is the Passed Percentage of the Candidates for a particular category. We define various terms in their codes as below:

a) Category

Category	Code
General	1
MOBC	2
OBC	3
SC	4
ST	5

b) Performance

Pass Percentage	Performance
>= 90%	1
>=75%	2
>=60%	3
>=45%	4
< 45%	5

c) Location

Location	Code
urban areas colleges	0
rural areas colleges	1

d) Gender

Gender	Code
Male Candidates	0
Female Candidates	1

The meaning of the data fields as depicted in the sample Table 2 are same Table 1 except one field i.e. 'Gender'. Now the stage is set and ready to perform.

B. Experiments performed by Rattle

The main objective in this paper is to select the best suited models for performing the statistical analysis of the datasets. We used one Xeon based Database Server for the experiments. The rattle package was used for the same. The data is imported to R which was stored in .csv format. The target data was categorical data and the partition chosen was 70/30/0. If one explores the data, one may visualise the data by using box plot, histogram, cumulative and benford curves. The histogram, the cumulative and benford curves are presented in the figures I,II,III and IV. Now, one may use the Model tab and select all the models for the comparison. The models are of type tree, random forest, boost, support vector machine, regression models and neural network. The data is evaluated through all the models. Our goal is to find the best suited models for the data through ROC curve.

C. Evaluation of the Experiments

In the figure V, we have placed one of the ROC curves for the category data. The followings are the actual findings using the Rattle based on the category wise data.

Area under the ROC curve for the rpart model on categoryba.csv [validate] is 0.8814

Rattle timestamp: 2014-05-06 06:48:54 sadiq

==========

Area under the ROC curve for the ada model on categoryba.csv [validate] is 0.9425

Rattle timestamp: 2014-05-06 06:48:55 sadiq

==========

Area under the ROC curve for the rf model on categoryba.csv [validate] is 0.9221

Rattle timestamp: 2014-05-06 06:48:55 sadiq

==========

Area under the ROC curve for the ksvm model on categoryba.csv [validate] is 0.9301

Rattle timestamp: 2014-05-06 06:48:55 sadiq

==========

Area under the ROC curve for the glm model on categoryba.csv [validate] is 0.8980

Rattle timestamp: 2014-05-06 06:48:55 sadiq

==========

Area under the ROC curve for the nnet model on categoryba.csv [validate] is 0.7393

Rattle timestamp: 2014-05-06 06:48:55 sadiq

From the above ROC curve analysis, it is quite clear that whose area under ROC curve for a particular model is 1 or close to 1, that model is best suited for that data. The Statisticians can further analyze the data based on that model.

The models best suited for our category-wise data are ada model (with area value is 0.9425), rf model (0.9221), ksvm model (0.9301).

If we generate the ROC curve for the gender specific data, then we find the following:

Area under the ROC curve for the rpart model on Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

==========

Area under the ROC curve for the ada model on Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

==========

Area under the ROC curve for the rf model on Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

==========

Area under the ROC curve for the ksvm model on Gender_BA.CSV [validate] is 0.9982

Rattle timestamp: 2014-05-07 19:55:54 sadiq

==========

Area under the ROC curve for the glm model on Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:54 sadiq

==========

Area under the ROC curve for the nnet model on Gender_BA.CSV [validate] is 0.9999

Rattle timestamp: 2014-05-07 19:55:54 sadiq

We may conclude from the above that almost all the models are would deliver better results, but rpart, ada, rf and glm models are best suited.

IV. Conclusions and Future Work

The Rattle package provides a GUI platform toward using R as a programming language. Rattle is open source data mining tools packed under the regime of R. In this paper, two data sets were mined. If one compares the two data sets results, then it may be concluded that ada, rf models are best suited for the data that were mined. We hance found that the female candidates of the University did better than the boys' candidates and the rural candidates did better performance than the urban candidates'(Refer to the figures below). Moreover, as this paper dealt with only one examination i.e. Bachelor of Arts, there are lots of another Examinations to deal with as well as one may extract valuable patterns and information from them. The future plan is to compare entry and exit data of TDC students of different colleges affiliated to Dibrugarh University.

V. Acknowledgments

The authors express their gratefulness to Prof. Alak Kr. Buragohain, Vice-Chancellor, Dibrugarh University for his inspiring words and allowing them to use the Examination data of the University. They generously thank Mr. N.A. Naik, Senior Programmer, Mumbai based firm for helping us to extract the .csv files from the SQL Server database.The authors would like to offer gratitude to Prof. Jiten Hazarika Department of Statistics, Dibrugarh University for his valuable ideas.

References

[1] The Dibrugarh University website: www.dibru.ac.in

[2] John Silltow, August 2006 : Data Mining 101: Tools and Techniques, http://www.internalauditoronline.org/

[3] Witten, I.H. and Frank, E. 1999. Data Mining:Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman, San Francisco, CA.

[4] Baker, R.S.J.d.: Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (eds.) To appear in International Encyclopedia of Education, 3rd edn. Elsevier, Oxford (2010)

[5] Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.

[6] Graham Williams, Rattle: A Data Mining GUI for R, The R Journal Vol. 1/2, December 2009 ISSN 2073-4859.

TABLE I. Sample Data for Year-Wise College-Wise Category-Wise Location-Wise Data of The B.A. Candidates

Year	CollegeCode	Category	Appeared	Passed	PassPercentage	Performance	Location
2010	103	1	2	2	100	1	1
2010	103	2	3	3	100	1	1
2010	103	3	25	25	100	1	1
2010	103	4	4	4	100	1	1
2010	103	5	11	8	72.73	3	1

TABLE II. SAMPLE DATA FOR YEAR-WISE COLLEGE-WISE GENDER-WISE DATA OF THE B.A. CANDIDATES

Year	CollegeCode	Gender	Appeared	Passed	PassPercentage	Performance
2010	101	0	46	42	91.3	1
2010	101	1	57	51	89.47	1
2010	102	0	57	47	82.46	1
2010	102	1	66	58	87.88	1

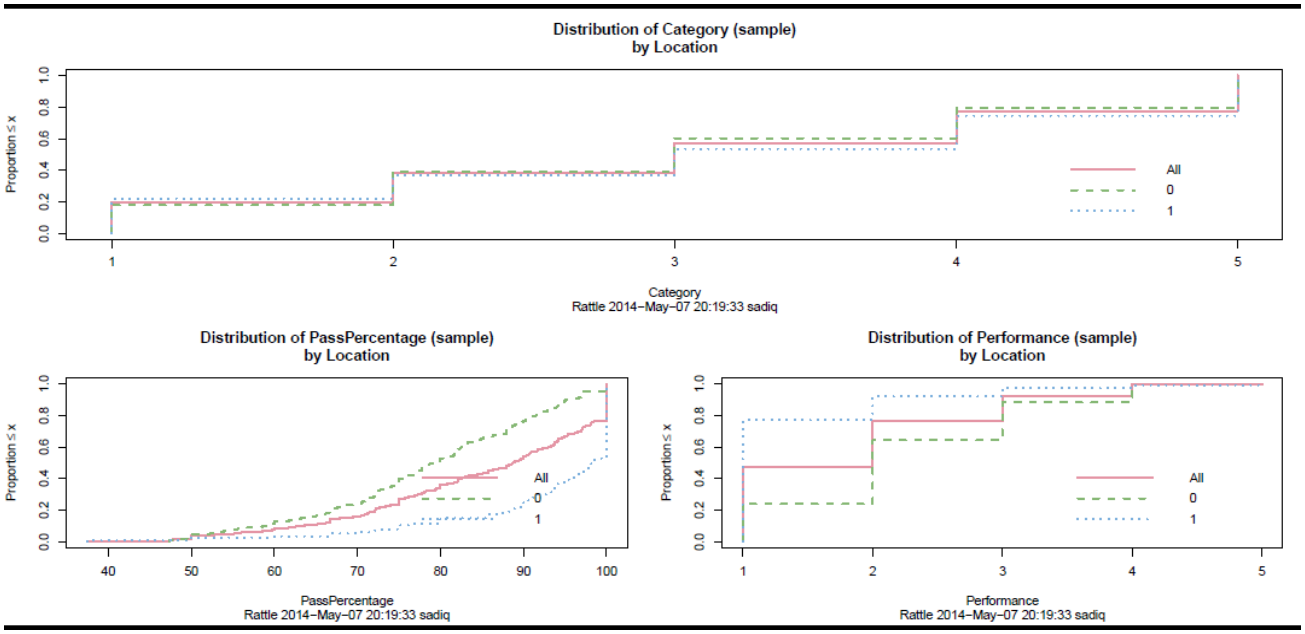


Fig. 1. Cumulative Diagram showing category-wise, Pass Percentage-wise, Performance-wise distribution on the basis of Location

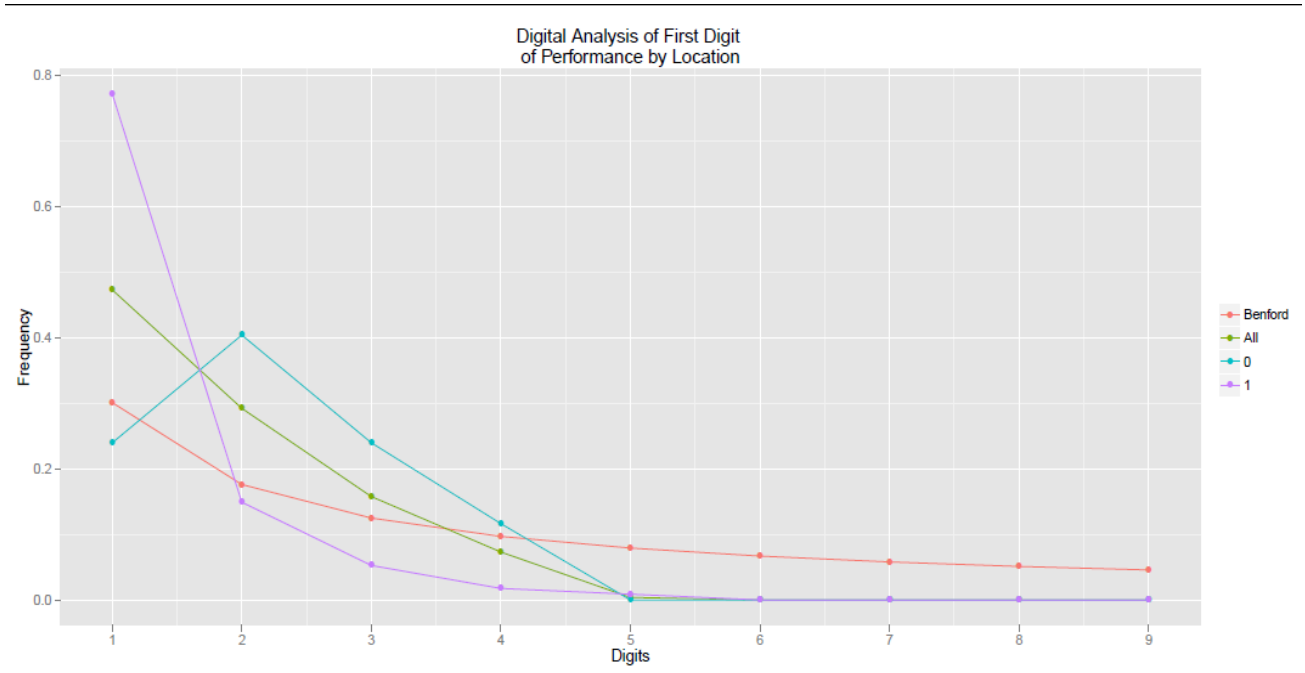


Fig. 2. Benford Diagram showing the performance by Location of the Candidates.

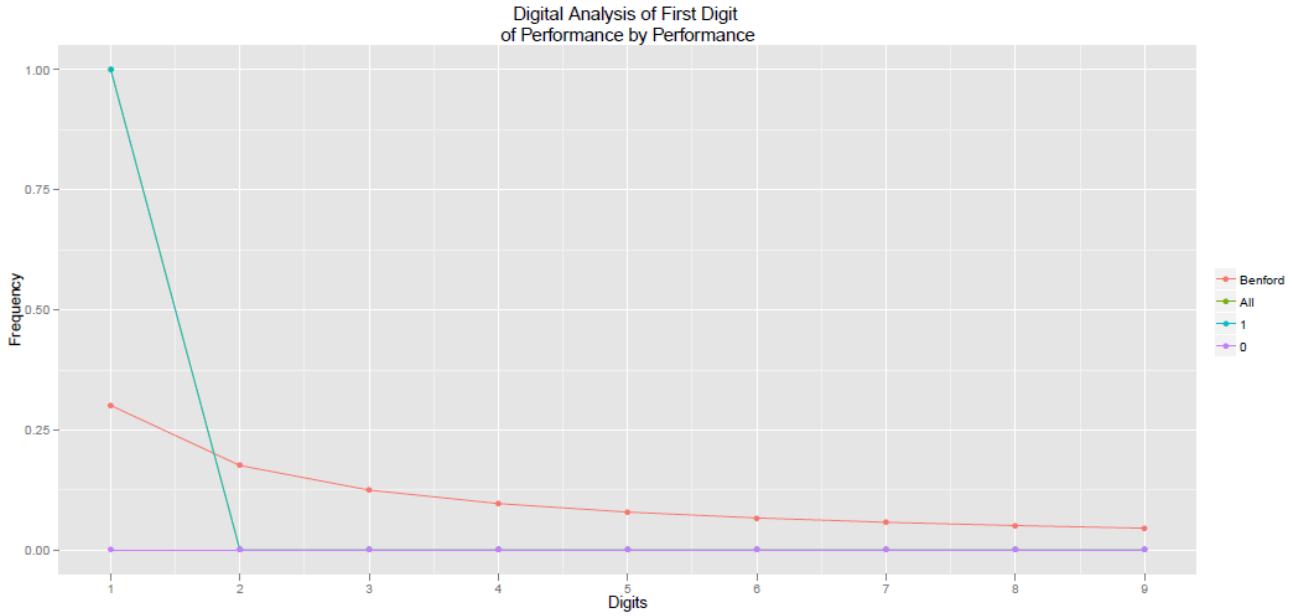


Fig. 3. Benford Diagram showing the performance by Gender of the Candidates.

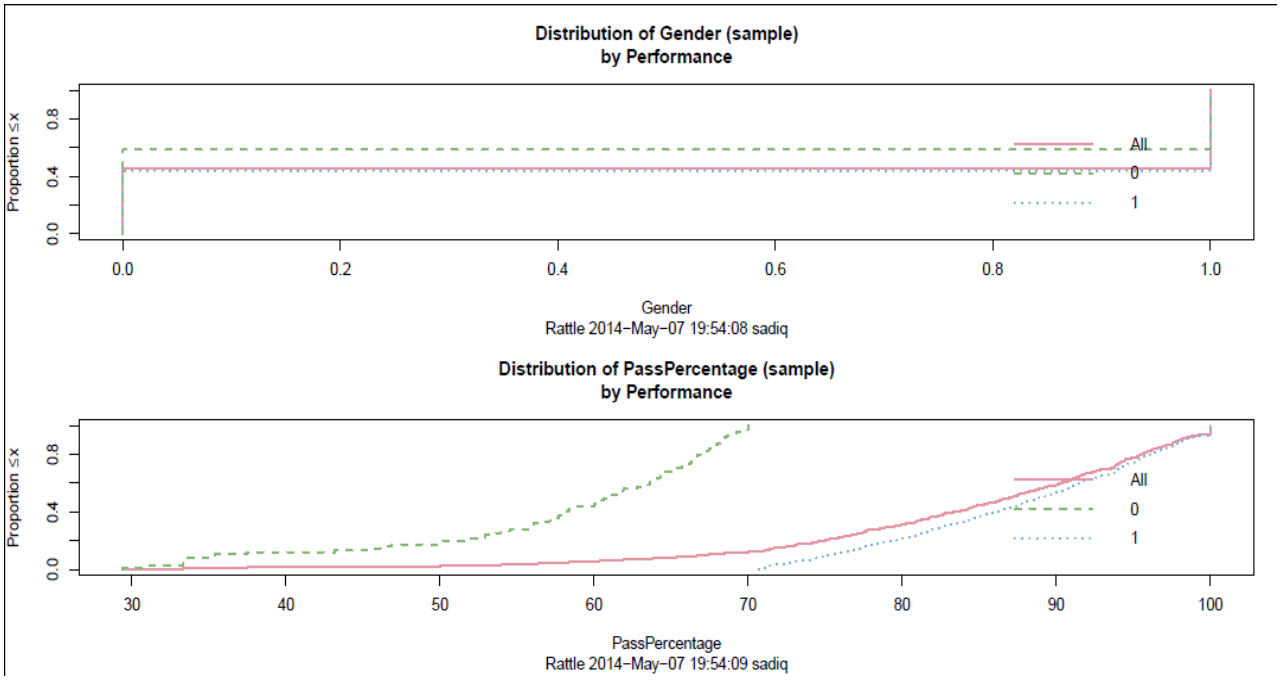


Fig. 4. Cumulative Diagram showing the performance by Pass Percentage and Gender wise.

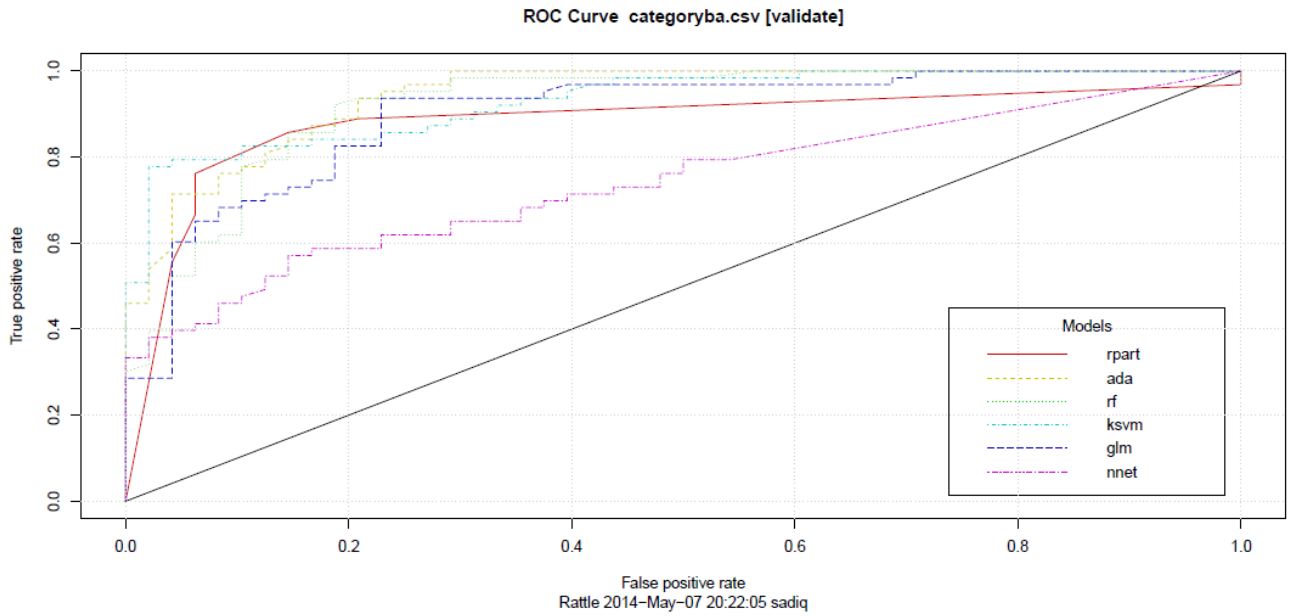


Fig. 5. ROC Curve for the first experiment i.e. performance by category.

Individual Syllabus for Personalized Learner-Centric E-Courses in E-Learning and M-Learning

Khaled Nasser ElSayed
Computer Science Department, Umm Al-Qura University

***Abstract*—Most of e-learning and m-learning systems are course-centric. These systems provided services that concentrated on course material and pedagogical. They did not take into account varieties of student levels, skills, interests or preferences. This paper provides a design of an approach for personalized and self-adapted agent-based learning systems for enhancing e-learning and mobile learning (m-learning) services to be learner-centric. It presents a modeling of goals of different learners of a corporate training in computer courses in an educational institute. It figures how to customize and personalize learning paths (course syllabus) for e-learning and m-learning platforms. The delivering of e-courses become personalized learner-centric, which improves learning outcome, satisfaction of learners and enhances education.**

Keywords*—AI; Agent; education; e-Learning; m-Learning; Semantic Net

I. Introduction

E-learning is nowadays recognized as one of the efficient methods to respond to the requirements of open and distance learning. In the e-learning system, several traditional learning styles should be combined with the learner-centered approach. It needs a good notation to represent the requirements of the e-learning system [1].

In the dynamic changes information environment without prior modeling, it can independently plan complex operation steps to solve practical problems, can independently discover and obtain the available resources the learners needed and then provide the corresponding services under the circumstance that the learners do not take part in [2].

Intelligent agents are task-oriented software components that have the ability to act intelligently. They may contain more knowledge about the needs, preferences and pattern of the behaviors of a person or a process as in [3].

The agent has to collect users' personal interests and give fast response according to the pre-specified demands of users. The personal agent can discover users' personal interests voluntarily without bothering the users. It is very suitable for personalized e-learning by voluntarily recommending learning materials [4].

Intelligent agents should have the ability of adaptive reasoning. They must have the capability to access information from other sources or agents and perform actions leading to the completion of some task. Also, they must control over their internal state and behavior and work together to perform useful and complex tasks. Thus, they should be able to examine the external environment and the success of previous actions taken under similar conditions and adapt their actions [5].

Educators, using Web-based learning environments, are in desperate need for non-instructive and automatic ways to get objective feedback from learner in order to better follow the learning process and appraise the online course structure effectiveness. On the learner side, it would be very useful if the system could automatically guide the learner's activities and intelligently recommend online activities and resources that would favour and improve the learning. The automatic recommendation could be based on the instructor's intended sequence of navigation in the course material, or, more interestingly, based on navigation patterns, of other successful learners [6].

A large proportion of university students are now part of the millennial generation. Mobile technology is now an integral part of their everyday life. The most educational use of mobiles by university students are calculator usage, text messaging, and English dictionary. Having a mobile with multiple capabilities, long battery life and good network coverage are the most influential factors in the educational use of mobiles [7].

The proposed design will bring learner-centric tailoring of materials according to the interest and goal of the learner. In this system, each time, objectives of a learner are changed, his category(Class of learner) is updated. Then, a new individual syllabus is created for building of a tailored and personalized learner-centric e-course. Learner can access e-services, anytime, anywhere, from any PC, laptop, tablet computer, pocket PC or any GSM mobile phones.

Section II will navigate through systems of e-learning and m-learning and gives a comparisons between them, while section III, will describe the structure of multi agents and system semantic net knowledge base. Section IV will explain the process of building a centric a customized e-course, while, section V will navigate through the process of creating individual syllabus for a special e-course . Finally, section VI will give a system conclusion and predicted future work.

II. E-Learning and M-Learning Systems

A. e-Learning vs. m-Learning

E-learning emerges as a solution to conventional learning methods. It has turned out that the learning process can significantly be improved if the learning content is specifically adapted to individual learners' preferences, learning progress and needs. An agent in e-learning application is situated in the

learning environment and performs the pedagogical tasks autonomously [8].

Most of the traditional e-learning systems are not learner-centric, and they often ignore the diversity of learner population; thus very often their service is not able to directly or effectively match the learner's goal [9].

Since the students and teachers are on different time and spare in an e-learning environment, the learning status of a student is difficult to be controlled by teachers. In current learning platforms, they neither analyze the causes of learning inefficiency of users, nor generate new learning material and testing. The former keeps the learners from not using these learning systems anymore because they are confusing; the latter leads to out-of-date materials and the learners could not get any new knowledge[10].

Mobile learning as a kind of learning model allowing learners to obtain learning materials anywhere and anytime using mobile technologies and the Internet. It is necessary that the elements of mobile learning are organized correctly and the interactions between the various elements are combined in an efficient and optimum way so that the mobile learning is successful and the implementation is efficient [11].

On the other hand mobile learning decreases the restrictions of learning environments by creating more flexibility, focusing on mobile technology and the mobility of learning environment. Therefore, the mobile learning is always concerned for its availability to different learning materials. Meanwhile this kind of learning is completely interactive and pleasurable and it simply creates more effective and amusable learning. The mobile learning is a developed kind of electronic learning that in relation to the other kinds of electronic learning provides learners with more facility to access the learning contents. It is evident that the mobile learning brings a communicative and interactive property for users. M-Learning is a kind of e-learning through mobile devices. Mobile learning is a compound facility that includes two fields: the computer aided mobiles and the Electronic learning [12].

With the support of today's mobile technologies to e-learning within d-learning (distance learning) concept, the notion of m-learning provided technological progress in education [13]. Saadiah and et al made a comparison of learning paradigms in [14]. Part of this comparison is shown Table 1.

B. *Samples of e-Learning and m-Learning Systems*

There are too much work done in the field of e-learning and e-teaching based on agent. Gascuena and Fernadez-Caballeroe introduced in [15] an Agent-based Intelligent Tutoring System for enhancing E-Learning/E-Teaching, where agents monitor the progress of the students and propose new tasks. De Antonio presented in [16] an architecture of intelligent virtual environment based on agent technology. Also, a similar one for nurse training is offered in [17]. Tang offered the implementation of a multi-agent intelligent tutoring system for learning the programming languages [18]. According to Java Agent for distance education (JADE) frame work, Silveira and Vicari carried out their system Electrotutor which is Electrodynamics distance teaching environment [19].

TABLE I. COMPARISON OF LEARNING PARADIGMS.

Criteria	e-learning	m-learning
Concept	Learn at the right time	Learn at the right place and time
Permanency	Learners can lose their work.	Learners may lose their work. Changes in learning devices or learning in moving will interrupt learning activities
Accessibility	System access via computer network	System access via wireless networks
Immediacy	Learners cannot get information immediately	Learners get information immediately in fixed environments with specified mobile learning devices
Interactivity	Learners' interaction is limited	Learners can interact \with peers, teachers, and experts in specified learning environment
Context-Awareness	The system cannot sense the learner's environment	The system understands the learner's situation by accessing the database

ElSayed proposed in [20] a multi-agent system that could get learner profile knowledge at his logging to the e-course. Then system can help users and advises them in their on line learning. It advised learners for better navigation through e-course contents by offering some links or jumping over course resources.

El Bouhdidi and et al., proposed in [21] a model of E-Learning based on a process of coupling of ontologies and multi-agent systems for a synergy of their strengths. Indeed, this model allows human agents (students, teachers and instructional designers) to cooperate with software agents to automatically build courses guided by relevant learning objectives. In addition, it allows learner to follow his training at his own space and according to his preferences, either individually or jointly with others (students or tutors).

Nordin proposed in [22] a conceptual framework for mobile learning applications that provides systematic support for mobile lifelong learning experience design. It concerns four perspectives: generic mobile environment issues, learning contexts, learning experiences and learning objectives. The paper also explores crucial factors and design requirements for the mobile learning environment. It also suggests how mobile learning applications can be designed with an understanding of these factors and requirements and further applied to lifelong learning.

III. CENTRIC-LEARNING SYSTEM

A. *Multi-Agent in Cenric-Learning Sysetm*

The multi-agent systems (MAS) are a society organized, constituted by semi autonomous agents, which interact with others, aiming to resolve collaboratively some problems, or to achieve some individuals or collectives goals. The agents may be homogeneous or heterogeneous and have common goals or

not, but still maintain a degree of communication between them [23].

The proposed system is an upgrading of the system in [20] from learner advising task to adaptive and learner-centric task. This is done to reason with learner requests and wishes, and target. It includes new additional agents. Its agents, shown in Fig. 1, can be explained as follows :

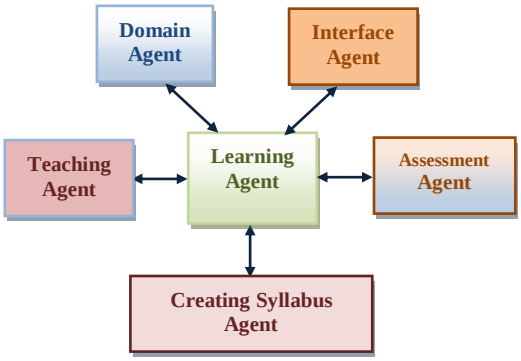


Fig. 1. Multi-Agent in the System

- Interface Agent (Iagent) is responsible of interaction with the learner to acquires his account personal information and profile knowledge. Then, it stores them in the learning profile Knowledge Database (KDB). Also, it consults the materials of tailored courses and assessments to the learner.
- *Learning Agent* (Lagent) is an important control agent, that is responsible of many tasks including managing the learning process. It controls all other agent in the system, initiate their work and collect their gains. Also, It Analyses profile knowledge of learners, and updates their profile record. As example, it receives assessment results from Aagent and evaluate learning efficiency of learner and update the learning KDB.
- Creating Syllabus Agent (CSagent) is the main player in the proposed system. It receives the profile knowledge of the learner and got materials and their pre-requisite, and finds out a classification for that learner. Finally, it generates a customized individual course syllabus relevant to learner classification.
- *Domain Agent* (Dagent) receives a learning path from Lagent and locates topics specified in the individual syllabus in the server of the educational institute. Finally, it creates a tailored individual e-course, customized for certain learner or a group (class) of learners.
- *Assessment Agent* (Aagent) which is an external agent system for creating an assessment (quiz or test) in [24]. It receives a request from Lagent to build an assessment to be conducted to the learner, under some conditions. This agent selects exercise or questions randomly to creates quizzes or tests with two level of difficulties for each topic(s) from the course material. It also grades the assessment and give the correct answers for each question.
- *Teaching Agent* (Tagent) retrieves the prerequisite of each topic or page in the course material page. It reviews if a course material is suitable to certain learner currently or not.

B. Object-Oriented Semantic Network

An object-oriented semantic net is designed to the proposed system. It include classes for Category of learner (class based on learner objectives), Case (represent a student to be classified), topic (main topics and sub topics), objective (main-objectives and sub-objective, syllabus (current learner syllabus, syllabus for Category, and syllabus for general courses). All of those classes are represented in nodes related to each other with links. Links represent relation between those classes as shown in Fig. 2.

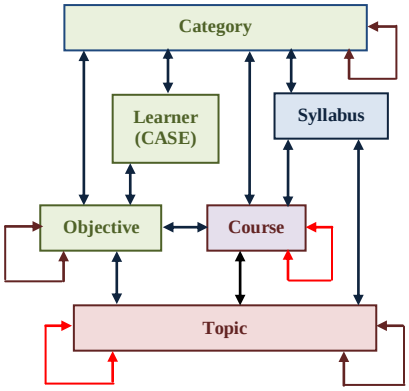


Fig. 2. Structure of Semantic Net of the System

Links are so important as usual in the semantic net. They might relate a node to itself (to represent a tree of objects related to that node from main topic to sub topic) as for category, objective, course, and topic. But topic class has another link with itself, to represent a link between certain topic and its prerequisite topic. Also, they can relate an object related to a node to another object from different node.

Training materials are prepared as isolated topics. Each topic is stored in a separate file, and has a class in the semantic net to store knowledge about it, as shown in Fig. 3.

Topic	Course	Perquisite Topic	Sub Objective	Question Bank

Fig. 3. Structures of Topic Class.

Each topic has its questions stored in a question bank to be used in assessment creations. Each topic has a sub objective, while each course has a main objective. All are consulted to learner o select his objectives from a list of objectives of all courses and topics in the server.

IV. Building Learner-Cenric E-Course

The suggested approach can automatically customize corporate training courses depending on learner objectives. It builds training courses with relative to knowledge level and skills (past), preferences (present), learning performance, and objectives (future). Fig. 4, presents the algorithm of building a personalized tailored e-course.

1. **READ learner Profile to get his objectives and skills.**
2. **IF the learner has a main objective**
   - **MOVE course name** of training course-of the same objective- **to variable CN.**
   - **FIND the Perquisite Course CNPRE.**
   - **CHECK learner profile,**
   - **IF learner finished CNPRE**
     MOVE the syllabus of CN course to the learner Syllabus
     ELSE MOVE the syllabus of CNPRE course to the Learner syllabus, GO to 6.
3. **ELSE GET all sub-objectives of the learner as attributes of a CASE.**
4. **IF there is a CLASS for CASE**
   - **MOVE syllabus of the CLASS to Learner Syllabus, GO to 6.**
5. **ELSE CALL CREATE Individual Syllabus**
6. **// creating tailored e-Course using the Individual Syllabus.**
   FOR each topic in the Syllabus table DO
   - **CONSULTE Topic pages to learner.**
   - **CONSULT an assessment in Topic Material to learner.**
   - **EVALUATE Learner Answers.**
   - **IF learner pass the assessment**
     UPDATE learner Profile,
     GO to next Topic in the Syllabus
   - **ELSE REPEATE Topic Again.**
7. **//After Navigating all topics in the Syllabus table**
   - **CONSULT an assessment in Course Material to learner.**
   - **EVALUTE Learner Answers.**
   - **IF learner pass the assessment**
     UPDATE learner Profile, Go to 1.

Fig. 4. Algorithm for tailoring an e-course

After the learner specified his objectives by selecting from consulted list, the system uses pre assessments to evaluate student knowledge before stating a new e-course. Also, at succeeding in assessment of an e-course, it evaluates learner by post an assessment. In both situations, it updates learner profile.

For each learner and similar (class of learner) there is a suitable learning path. The system classifies each learner according to his profile knowledge. It assigns him to a suitable class, if there is. If there is no suitable category, the approach will create new one. Class of learner is reviewed and updating after finishing any learning path and assessment evaluation.

V. Creating Individual Syllabus

When the system couldn't find a ready syllabus for the current learner, it calls an algorithm to create a new a tailored syllabus (customized learning path), specially for that learner. Fig. 5, present the entry structure of the requested individual syllabus table. While, Fig. 6, present the algorithm suggested for creating the individual syllabus for the current learner and next similar ones.

Topic Order	Topic Code	Perquisite Code

Fig. 5. Structure of Individual Syllabus Table

1. **FOR each sub-objective in learner objectives list DO**
   - **GET** topic(s) linked to objective.
   - **CREATE** new entry in the table of Individual Syllabus for each topic.
   - **INSERT** Topic Code (TC) and its Prerequisite Code (PC) in the new entry for each topic.
2. **FOR each entry(i) in the Syllabus table DO**
   - **READ** current PC_i from Syllabus entry.
   - **CHECK** learner profile for the prerequisite
   - **IF** learner didn't study the prerequisite
     a. **LOOK** up for an entry (j) in Syllabus table, where $TC_j = PC_i$.
     b. **IF** found, interchange the two entries (if $j > i$).
   - **ELSE**
     - **GET** new entry for the requested prerequisite Entry from the table of material topic.
     - **INSERT** the new entry in order i-1.
3. **CREATE** Category (CLASS) with the objectives of the current learner.
4. **UPDATE** semantic net links.
5. **RETURN with** the individual learner Syllabus.

Fig. 6. Algorithm for creating individual syllabus

VI. Conclusion

The presented paper provided a design for an approach used for personalized and learner-centric agent-base systems for enhancing e-learning and m-learning. The proposed system used to satisfy different wishes of learners of a corporate training in computer and English courses. It could customize and personalize learning paths by creating individual syllabus

for an e-course. It could build a tailored e-course for training in e-learning and m-learning platforms. The delivering of an e-courses became according to learner learner-centric, which improves learning outcome objectives, skills, and experience which resulted in satisfaction of learners and enhancing education. In future, this design will be improved to create more flexible individual syllabus that can be updated while learner is studying his special e-course, to meet his learning capabilities. Also, it will be upgraded to be applied over ubiquitous learning (u-learning) and using ubiquitous computing environment.

REFERENCES

[1] Zhi Liu & Bo Chen, "Model and Implement an Agent Oriented E-Learning System", Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05), 2005.

[2] Ying-Han Fang1 and Lei Shao2, "Agent-Based E-Learning System Research and Implementation", proceedings of the 7th International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008, 4080-4084.

[3] H. K. Mohammed, "An Intelligent Agent to Teach C-Language", proceedings of ICECS'97, Cairo, Egypt, December 15-18, 1997, 483-488.

[4] Jin-Ling Lin and Ming-Hung Chen, "An Intelligent Agent for Personalized E-Learning", 8th International Conference on Intelligent Systems Design and Applications - Volume 01,27-31,2008.

[5] M. Nissen, "Intelligent Agents: A Technology and Business Application Analysis", Telecommunications and Distributed Processing, November 1994.

[6] O. R. Zarane, "Building a Recommender Agent for e-Learning Systems", proceedings of the international Conference on Computers in Education(ICCE'02), pp 55-59, Dec. 2002.

[7] Z. Taleb, A. Sohrabi, "Learning on the move: the use of mobile technology to support learning for university students", International Conference on Education and Educational Psychology (ICEEPSY 2012), Procedia - Social and Behavioral Sciences 69, pp. 1102 – 1109, 2012.

[8] N. Sivakumar1, K. Vivekanandan2, B. Arthi3, S.Sandhya4, Veenas Katta5, "Incorporating Agent Technology for Enhancing the effectiveness of E-learning System", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 3,May 2011.

[9] D. Li, Z. Shen, Y. Miao, C. Miao, and R. Gay, " A Goal Oriented e-Learning Agent System", KES 2005, LNAI 3681, pp. 664-670, 2005.

[10] HL.Tsai, CJ. Lee, WH. HSU, and YH. Chang, "An Adaptive E-learning System based on Intelligent Agents", Proceeding of the 11th International Conference on Applied Computer and Applied Computer Science, WSEAS.US, pp 139-142, Steven point, Wisconsigin, April 2012.

[11] F. Ozdamli, N. Cavus, " Basic elements and characteristics of mobile learning", Procedia - Social and Behavioral Sciences 28, pp. 937-942, 2011.

[12] M. Akhshabi, J. Khalatbari, M. Akhshabi, "An experiment on conducting mobile learning activities on the virtual university", Procedia - Social and Behavioral Sciences 28 , pp. 384 – 389, 2011.

[13] T. Korucu and A. Alkan, " Differences between m-learning (mobile learning) and e-learning, basic terminology and usage of m-learning in education", Procedia Social and Behavioral Sciences 15, pp. 1925–1930, 2011.

[14] S.Yahya, E.A. Ahmad and K. Abd Jalil, "The definition and characteristics of ubiquitous learning: A discussion", International Journal of Education and Development using Information and Communication Technology (IJEDICT), Vol. 6, Issue 1, pp. 117-127, 2010.

[15] J.M. Gascuena and A. Fernandez-Caballero, "An Agent-based Intelligent Tutoring System for Enhancing E-Leaning/E-Teaching", International Journal of instructional Technology and Distance Education, itdl.org, vol. 2, No. 11, pp 15-26, Nov. 2005.

[16] A. de Antonio, and et al., "Intelligent Virtual Environments for Training:An Agent-based Approach. 4th International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS'05). Hungary, 15-17, 2005.

[17] M. Hospers, E. Kroezen, A. Nijholt, H.J.A. op den Akker, and D. Heylen, "An agent-based intelligent tutoring system for nurse education", Application of Intelligent Agents in Health Care, J. Nealon and A. Moreno (eds), 143-159, 2003.

[18] T. Y. Tang and A. Wu, "The implementation of a multi-agent intelligent tutoring system for the learning of a computer programming", Proceedings of 16th IFIP World Computer Congress-International Conference on Educational Uses of Communication and Information Technology, ICEUT, 2000.

[19] R.A. Silveira and R.M. Vicari, "Developing Distributed Intelligent Learning Environment with JADE –Java Agents for Distance Education Framework", International Conference on Intelligent Tutoring Systems, LNCS, 2363, 105-118, ITS 2002.

[20] Khaled Nasser ElSayed, "A Multi_Agent Advisor System for Maximizing e-Learning of an e-Course", International Journal of Advanced Research in Artificial Intelligences(IJARAI), Vol. 3, No.5, May 2014.

[21] J. El Bouhdidi, M. Ghailani, O.Abdoun, and A.Fennann, "A New Approach based on a Multi-ontologies and Multi-agents System to Generate Customized Learning Paths in an E-Learning Platform", International Journal of Computer Applications (0975 –8887), Vol. 12– No.1, December 2010.

[22] N.Nordin, M. A. Embi, and M. Md.Yunus, " Mobile Learning Framework for Lifelong Learning", International Conference on Learner Diversity, 130-138, 2010.

[23] A. BENNANE, "Tutoring and Multi-Agent Systems: Modeling from Experiences", Informatics in Education, , Vol. 9, No. 2, 171–184, 2010

[24] Khaled.Nasser ElSayed," A Tool for Creating Exams Automatically From an OO Knowledge Base Question Bank", International Journal of Information and Education Technology, IJIET Vol. 3, No.1, 27-31, Feb. 2013.

AUTHOR PROFILES

The Author is Dr. Eng. Khaled Nasser. ElSayed. He was born in Cairo, Egypt 9 Oct. 1963. He have got his PhD of computers and systems from Faculty of Engineering, Ain Shams University, Cairo, Egypt, 1996.

He has worked as an associate professor of computer science, in Umm-AlQura Uni. in Makkah, Saudi Arabia since 2008. Artificial Intelligence is his major. His interest research is Distant Education, E-Learning, and Agent. Dr. Khaled Nasser ElSayed translated the 4th edition of "Fundamentals of Database Systems", Ramez Elmasei and Shamkant B. Navathe, Addison Wesley, fourth edition, 2004, published by King Saud University, Riyadh, Saudi Arabia, 2009. He is also the author several books in programming in C & C++, Data Structures in C& C++, Computer and E-Society, Database Design and Artificial Intelligence.

Security Policies for Securing Cloud Databases

Ingrid A. Buckley
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

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

***Abstract*—Databases are an important and almost mandatory means for storing information for later use. Databases require effective security to protect the information stored within them. In particular access control measures are especially important for cloud databases, because they can be accessed from anywhere in the world at any time via the Internet. The internet has provided a plethora of advantages by increasing accessibility to various services, education, information and communication. The internet also presents challenges and disadvantages, which include securing services, information and communication. Naturally, the internet is being used for good but also to carry out malicious attacks on cloud databases. In this paper we discuss approaches and techniques to protect cloud databases, including security policies which can realized as security patterns.**

Keywords—relational database; cloud; security; threats; hackers, security patterns; cloud database

I. Introduction

Technology has changed the way businesses conduct their daily tasks and processes. Most businesses have evolved in how they utilize data, most times they have to collect, query, manipulate and store data rapidly in order to provide services to their consumers. Databases are one of the most common resources used for business. Today, Relational Database Management System (RDBMS) is a staple resource in businesses of all types. In particular, cloud databases provide increased accessibility to valuable information that is stored to carry out business functions. The main advantages of the cloud are increased availability, scalability, elasticity and performance of databases.

The Internet, since its inception, has been continuously evolving, creating both problems and solutions. The cloud lives in the internet and has inherited the benefits, challenges and problems associated with the Internet. The cloud is still a relatively new approach in how technology and resources are shared through a network: the Internet. Cloud computing is by nature a dynamic and fast changing environment which is designed to provide services to various clients. The goals of these clients vary, from business owners, employees, customers to attackers. An attacker can take any form; this makes the job of security more complex and challenging. Since the introduction of cloud Databases, there has been sustained and increased attacks against web services and databases [1] which are primary aspects of the cloud. The goal of an attackers is to attack exploit the fundamental components of the cloud.

The paper is organized as follows. Section 2 presents background information on database security breaches. Section 3 provides an overview of cloud relational databases.

Section 4 presents some approaches to protect cloud databases. Section 5 discusses security patterns. Section 6 presents some related work. The paper concludes in section 7.

II. Security Breaches

On November 2013 the Target store databases were attacked, the personal and credit card information of 40 million customers was compromised [1]. In April 2010, hackers gained access to approximately 77 million PlayStation Network accounts. In this attack unencrypted credit card numbers, personal information and purchase history was compromised [2]. RSA servers were compromised my hackers which is the security division of EMC which is a huge storage company used by many financial institutions. EMC stores close 40 million authentication tokens used by employees to access corporate and government networks, the hackers were able to gain access to these tokens. Since this incident EMC has spent over 60 million to monitor the information of concerned clients. Similarly in August 2007 hackers attacked Monster.com and stole resume information of 1.3 million dollar job seekers [2]. These incidents are common and hackers continue to strengthen their efforts in attacking corporate, e-commerce and government systems. The problem associated with security breaches are far reaching and affect other aspect such as privacy and reliability. Security patterns can be used in software engineering/development solutions to solve security problems [3].

III. Cloud Relational Database Management System (RDBMS)

Database Management Systems (DBMS) provide an organized way to utilize data. They also secure information against system failure or tampering and permit data to be shared among multiple users. A Relational Database Management System (RDBMS) stores a collection of interrelated data that allows programs to access the data. A relational database allows the definition of data structures, storage, retrieval operations and integrity constraints. The data and relations between them are organized in tables. A table is a collection of records and each record in a table contains the same fields.

Fig. 1 illustrates a simplified example of the cloud architecture. Clients or users connect to the internet through their respective internet providers. Once the client is connected to the cloud, they have access to variety of infrastructure, services, and platforms. We are interested in The Relational

Database Management System (RDBMS) as shown in red below. The architecture of the cloud consists of virtual machines and hardware which consists of storage, servers and networks.

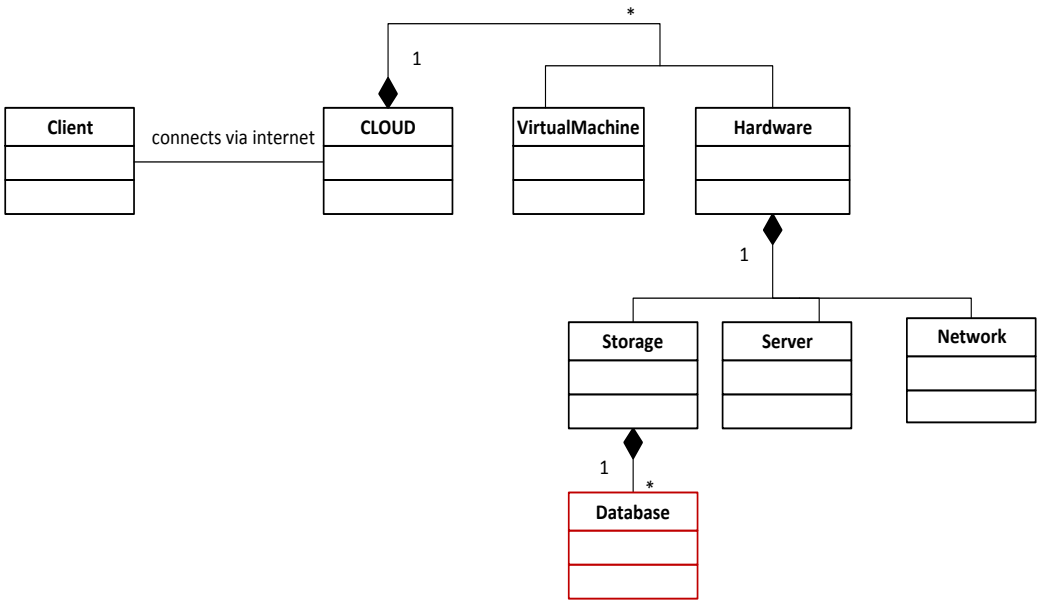


Fig. 1. Cloud Architecture

IV. Approaches and Techniques to Secure Cloud Databases

Security provides protection against unauthorized data disclosure, data modification, data destruction and denial of service. The database system is a fundamental aspect for security because it stores the persistent information, which constitutes most of the information assets for an institution. A private cloud is concerned with the internal needs of an organization. A public cloud sells resources to the general public. A hybrid cloud gathers resources from different clouds; it is a combination of public and private clouds. Due to the nature of the cloud environment, and the need to have active reliable security mechanisms, we enumerate here some common security and reliability policies which are used to protect cloud databases [8, 10]:

A. *Equations Security Policies:*

- **Least Privilege -** Limits access to resources by allowing only the minimal level of access, while still allowing an application to function normally.
- **Separation of Privileges** - Separates critical functions that can affect the security of the database into portions that is performed by different people or systems.
- **Authorization And Authentication -** Authorization defines permitted access to resources. Authentication verifies the identity of an entity requesting access to a resource [14].
- **Defense in Depth -** ensures that security controls are implemented at all levels of the information architecture including the database, network, sever, and operating system. This policy can be implemented using *Cloud Pools* [8].
    - A cloud pool is common set of resources that is shared by multiple tenants.
- **Logging and Auditing -** Tracks all activities by keeping a log of actions that may be relevant for security.
- **Information Hiding -** Conceals sensitive information with the use of cryptography and hashing functions.

B. *Reliability Policies:*

- **Redundancy** [13, 14] - The replication of critical components in a system or of a complete system with the intention of increasing the reliability of the system.
    - Additionally scheduling frequent backups of the DMBS is essential to restore corrupt/lost data if required.
- **Monitoring** [13, 14] - The constant checking of the state of a system to ensure that specifications are being met. This is a fundamental step because a security breach cannot be addressed if it is not detected.

Most databases do not implement all of the policies mentioned above; because this is not practical, due to the fact that increased security can result in a reduction in throughput and robustness. In particular, cloud databases have to respond quickly to requests in order to maintain their effectiveness. Many of these policies are described in pattern format to help developer's better implement security in cloud databases and environments.

Different security approaches and techniques have been proposed to secure databases that live in the internet or the cloud [10]. However despite the advances, hackers are still

finding ways to exploit vulnerabilities that go under the radar during development, testing and deployment. Access control is a very fundamental and critical security concern for databases that live online.

There are different types of users that interact with a database novice users, database administrators, programmers etc. Each of these requires a certain degree of leeway to perform their activities; as a result an authorized user can easily misuse their rights to compromise a database. Access control is generally achieved through one or more of the security policies discussed earlier.

V. Security Patterns

Patterns [3] embody the experience and knowledge of many designers and when properly catalogued, they provide a repository of solutions for useful problems. Initially used for improving code, patterns are becoming a staple tool to build secure systems [7, 9, 12]. The POSA [7] template defines a systematic way to describe patterns. It consist of approximately eleven units, each describes one aspect of a pattern. This template is designed to capture the experience and knowledge of professionals that have solved common problems. Patterns support best practices. Each unit of the POSA Template is described below:

a) ***Name*** - the name of the pattern should correspond to the generic name given to the specific type of attack in standard attack repositories such as CERT [11].

b) ***Intent or thumbnail description*** - A short summary of the intended purpose of the pattern, including which problem it solves

a. **Example** of a specific problem.

b. **Context** -this section describes where the pattern applies, including prerequisites and the general environment.

c. **Problem** - describe the forces which affect the solution, attacks.

d. **Solution** - describes the general idea of solving the problem, it includes UML models (static and dynamic), formalization.

e. **Implementation** – provides recommendations and hints for implementers

f. **Example resolved** - describes how the pattern solved the specific problem

g. **Known uses** - provides at least three examples of use in real systems

h. **Consequences** – provides advantages and disadvantages of the pattern's solution.

i. **Related patterns** - presents complementary or alternative patterns.

Security patterns describe mechanisms that control threats. Security patterns join the extensive knowledge accumulated about security with the structure provided by patterns to provide guidelines for secure system construction and evaluation. Security has had a long trajectory, resulting in a variety of approaches to analyze security problems and to design security mechanisms. It is helpful to capture this expertise in the form of patterns [7]. There are several books [10, 13] on security patterns and academic institutions that write and share security patterns. An attacker can attack a system from all levels. If a hacker does not find vulnerability in level n, then level n+1 or n-1 may have vulnerabilities that can be exploited. It is important to identify attacks at every level or stage in software development. Security patterns provide the following advantages:

- Security patterns embody experience and good design practices.
- They help to prevent errors, and save time.
- Can be reused.
- Provide guidelines to solve security problems.
- Provide best case solutions to common problems.

VI. Related Work

Google Cloud SQL [4] uses two level of access control before access is granted to the database. It first authorizes access to an instance using the host application ID or IP address. Second it authorizes the user or application to access the database. EDB Cloud database [5] provides role permission management, authentication of object permissions, auditing of user and application using logs and SQL injection protection. Oracle Database [6] provides label based access control to provide multi-level security and restricting access to data based on data classification and user security clearance. It also provides data encryption, data masking, blocks SQL injection attacks, and auditing of user and application activities.

VII. Conclusion

Database need effective access control security mechanisms to protect the data stored. In particular, cloud databases present a difficult problem because they can be accessed at anything through the Internet, therefore effective security mechanisms are necessary to protect them without affecting normal business operations. Not only is it important that a database as security controls but in addition, a wide variety of security policies are required at varying levels of a systems architecture to sufficiently protect it.

Acknowledgment

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

References

[1] CNN Money. (2013, December 19).Target: 40 million credit cards compromised. Retrieved from: http://money.cnn.com/2013/12/18/news/companies/target-credit-card/ Last Accessed: 1/30/2014.

[2] CNN Money. (2011, April). 9 of the worst security breaches ever. Retrieved from: http://money.cnn.com/galleries/2012/technology/1206/gallery.9-worst-security-breaches.fortune/2.html. Last Accessed: 1/30/2014.

[3] Ingrid. A. Buckley and Eduardo. B Fernandez. (2009).Three patterns for fault Tolerance. Proceedings of the OOPSLA MiniPLoP.

[4] Google Cloud.(2012, October). Levels of access control. Retrieved from: https://developers.google.com/cloud-sql/docs/access-control. Last Accessed: 2/28/2014.

[5] EDB Cloud database. (2005, May). Postgres Plus Advanced Server: Better protection for your critical data. Retrieved from: http://www.enterprisedb.com/docs/en/9.3/conncloud/Tutorial_Connecting_to_a_Cloud_Database_Cluster.htm. Last Accessed: 2/28/2014.

[6] Oracle Database 12^{c}. (2012, July). Security and Compliance. Retrieved from: http://www.oracle.com/technetwork/database/security/index.html. Last Accessed: 2/28/2014.

[7] M. Schumacher, E.B. Fernandez, D. Hybertson, F. Bushmann and P. Sommerland. (2006). Security Patterns: Integrating security and system engineering. West Sussex, England:Wiley.

[8] Google Cloud.(2012, October). Levels of access control. Retrieved from: https://developers.google.com/cloud-sql/docs/access-control. Last Accessed: 2/28/2014.

[9] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, M. Stal. (1996). *Pattern-Oriented Software Architecture: A System of Patterns*, vol. 1, Wiley.

[10] Oracle Corporation. (2012). Security in Private Database Clouds. Oracle White Paper. Retrieved from: http://www.oracle.com/technetwork/database/database-cloud/security-in-private-db-clouds-1733933.pdf

[11] CERT.(1988). Cybersecurity Engineering. Retrieved from: https://www.cert.org/about/.

[12] E. B. Fernandez.(2013). Security patterns in practice: Building secure architectures using software patterns., Wiley Series on Software Design Patterns.

[13] Ingrid. A. Buckley and Eduardo.B. Fernandez.(2011). Enumerating Software Failures to Build Dependable Distributed Applications. High-Assurance Systems Engineering (HASE), 2011 IEEE 13th International Symposium. 120 - 123. doi:10.1109/HASE.2011.35.

[14] I. Buckley and E.B.Fernandez, "Patterns Combing Reliability and Security", Procs. of the Third International Conferences on Pervasive Patterns and Applications, September 25-30, 2011, Rome, Italy.

Comparative Performance Analysis of Feature(S)-Classifier Combination for Devanagari Optical Character Recognition System

Jasbir Singh
Department of Computer Science
Punjabi University
Patiala, India

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University
Patiala, India

***Abstract*—this paper presents a comparative performance analysis of feature(s)-classifier combination for Devanagari optical character recognition system. For performance evaluation, three classifiers namely support vector machines, artificial neural networks and k-nearest neighbors, and seven feature extraction approaches viz. profile direction codes, transition, zoning, directional distance distribution, Gabor filter, discrete cosine transform and gradient features have been used. The first four features have been used jointly as statistical features. The performance has also been evaluated by using the combination of these feature extraction approaches. In addition, performance evaluation has also been done by varying the feature vector length of Gabor and DCT features. For training the classifiers, 7000 samples of first 70 classes (out of 942 classes), recognized in the earlier work have been used. Such a large number of classes are due to the horizontal and vertical fusion/overlapping characters. We have chosen first 70 classes as their percentage contribution out of 942 classes has found to be 96.69%. For testing, 1400 samples have been collected separately. A corpus of 25 books has been used for sample collection. Classifiers trained on different features, have been compared for performance evaluation. It has been found that support vector machines trained with Gradient features provide the classification correctness of 99.429%, and there is no significant increase in the performance with the increase in the feature vector length.**

Keywords*—Artificial Neural Network; DCT; Directional Distance Distribution; Feature extraction, Gabor; k-Nearest Neighbour; Profile direction codes; Support Vector Machines; Transition; Zoning

I. INTRODUCTION

Optical character recognition is a widely used technique for generating digital counterpart of printed or handwritten text. A lot of work has been done in this field, particularly from Devanagari script point of view. In one of the earlier work, Sinha and Mahabala [1] have used syntactic pattern analysis system with an embedded picture language for recognition of Devanagari script.

Bansal and Sinha[3,4], laid emphasis on the use of various knowledge sources at all levels in Devanagari document processing system. These knowledge sources are mostly statistical in nature. Chaudhuri and Pal [5] have suggested the primary grouping of characters, where each character is assigned to one of the three groups namely basic, modifier and compound character. A feature based tree classifier approach is then used for basic and modifier character recognition. As Devanagari script consists of several basic characters, half form of characters, vowel-modifiers and diacritics, therefore from character recognition point of view only 78 basic character classes are sufficient for the identification of these characters. But in Devanagari the characters fuse with each other and generate new compound characters which are very difficult to separate during segmentation phase of OCR process. These compound characters are commonly known as conjuncts.

Sinha and Bansal [2] have discussed the algorithms that can be used to segment the compound characters into its constituent symbols rather than treating the character as a single unit. But in our work we have considered these compound characters as single recognizable unit, so that the segmentation errors can be reduced. Kompalli, Nayak and Setlur [6] and Kompalli, Nayak and Govindaraju [7] have also discussed the wide range of challenges in Devanagari script including that of compound characters. These compound characters are the result of horizontal or vertical fusion of basic characters. As an example व + ् + य will form व्य and च + ् + च will result in च्च. From these two examples it is very much clear, that it is very difficult to decide from where to separate these compound characters into the constituent basic symbols and therefore treated as single recognizable unit.

The presence of conjuncts is not the only problem in segmentation but sometime the height of constituent characters in a word (to be segmented) is such that it causes segmentation problems, for example in the word समूह the height of constituent character ह is such that it can either lead to, over-segmentation of ह or under-segmentation of मू. Therefore we have also considered such character combinations (मू) as single recognizable units. It should be noted that if the constituent symbols are not connected, then they will be treated as separate recognizable units.

In order to identify the possible classes and their frequency of occurrence we [9] have used a corpus of approximately 3 million words, which comprises of Unicode data. We have identified 864 compound characters (apart from basic 78 characters), which comprise of both horizontally and vertically

fused characters, and consonants with lower vowel modifiers which makes a total of 942 recognizable units. As it is very difficult to handle such a large number of classes; therefore coverage analysis has been done to optimize the character class count. The analysis has been done on the basis of their frequency of occurrence. It has found that the first 70 classes contribute to 96.69% of the overall classes, as shown in Table I. Therefore in this work we have taken the first 70 classes to find an optimal Feature(s)-classifier combination. For evaluating the performance of feature(s)-classifier combinations 1400 test samples i.e. 20 samples per class has been collected separately.

TABLE I. PERCENTAGE CONTRIBUTION OF RECOGNIZABLE UNITS

Recognizable units	% contribution
20	82.0185
30	90.1112
40	93.4336
50	95.0826
70	96.6985

This paper is organized as follows. Section II describes the various classification techniques, like Support Vector Machines, Artificial Neural Network and k-Nearest Neighbor. Section III depicts the various feature extraction methods which are used in his work. In section IV the performance of all feature-classifier combinations has been presented, and in section V comparison of all combinations has been done. Conclusion and future scope is described in the section VI.

II. CLASSIFICATION METHODS

The task of classification is to assign an input pattern represented by feature vectors to one of many pre-specified classes. Here we have used three classifiers described here.

A. Support Vector Machines (SVM)

SVM's (Support Vector Machines) [10] are a useful technique for data classification. SVM is a supervised learning classifier. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (class label) and several attributes (features). The goal of SVM is to produce a model which predicts the target value. Given a training set of attributes-label pairs, (x_i, y_i), $i = 1, \dots, l$ where $x_i \in R^n$ and $y \in \{-1,1\}^l$, the support vector machines require the solution of the following optimization problem given by (1):

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \qquad (1)$$

$$\text{subjected to} \quad y_i \left(w^T \phi(x_i) + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Here training vectors x_i are mapped into higher dimensional space by the function ϕ. C >0 is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. Three kernel functions are listed below.

- Linear: $K(x_{i,} x_j) = x_i^T x_j$.
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0$.
- RBF: $K(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2), \quad \gamma > 0$.

Where $\gamma, r \text{ and } d$ are kernel parameters.

B. Artificial Neural Network (ANN)

Artificial neural networks are the computational models that consist of number of simple processing units called neurons distributed in layers namely input, hidden and output (Fig. 2) that communicate with one another over a large number of weighted connections. An artificial neural network is based on the operation of biological neural networks. The neurons in the ANN are the electronic counter part of the neurons of the human brain. Neuron of an artificial neural network consist of

- A set of input values (x_i) and associated weights (w_i)
- A function (φ) known as activation function that operates on the weighted sum (v_k evaluated by (2)), and maps the results to an output (y_k).

$$v_k = \sum_{j=1}^{p} w_j x_j \qquad (2)$$

The model in Fig.1 shows the interval activity of the neuron.

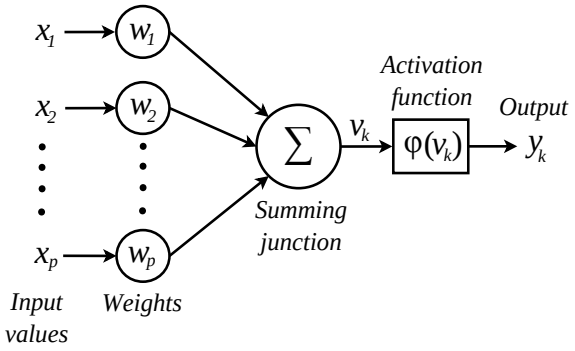


Fig. 1. A Neuron

Various activation functions (φ) available are: threshold, piecewise linear, sigmoid, Elliot and Gaussian etc. We have used Elliot and sigmoid as activation functions, which have been determined experimentally. There may be several hidden layers in the neural network, but we have used a single hidden layer. The number of hidden neurons is determined experimentally. A Neural network can be trained by using sample training data, and then the trained network can be used to predict the class of unknown test sample. Each neuron in output layer corresponds to each class.

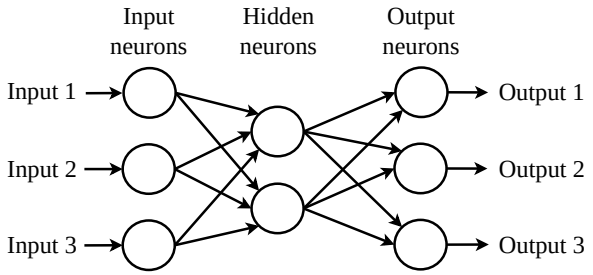


Fig. 2. A Neural Network

C. k-Nearest Neighbour (kNN)

The nearest-neighbor classifier is one of the simplest of all classifiers for predicting the class of the test sample. Training phase simply store every training sample, with its label. To make a prediction for a test sample, its distance to every training sample is computed. Then, keep the k closest training samples, where k≥1 is a fixed integer. Then a label is searched that is most common among these samples. This label is the prediction for this test sample.

This basic method is called the kNN algorithm. There are two major design choices to make: the value of k, and the distance function to use. We have chosen k = 1, 3, 5 and 7 and for the minimum distance, the metric employed is the Euclidean distance given by (3), which evaluates the distance d(x, y) between test and training sample.

$$d(x,y) = \| x - y \| = \sqrt{(x-y)*(x-y)} = \left(\sum_{i=1}^{m} (x_i - y_i)^2 \right)^{\frac{1}{2}} \quad (3)$$

Where $x, y \in \mathbb{R}^m$

III. Feature Extraction Methods

Feature extraction is the process of extracting distinctive information from the digitized sample. The aim of feature extraction is to describe the pattern by means of a minimum number of features or attributes that are effective in discriminating it among pattern classes.

A. Statistical Features

Statistical features describe a pattern in terms of a set of characteristic measurements extracted from the pattern. The statistical features which have been used are: Profile Direction Codes, Zoning, Transition and Directional Distance Distribution.

1) Profile Direction Codes: A variation of chain encoding has been used on left, right, top and bottom profiles. First the sample image is scaled to 50*50. For finding the left profile direction codes, the image is scanned from left, from top to bottom and local directions of the profile at each pixel are noted. Starting from current pixel, the pixel distance of the next pixel in east, south or west directions is noted. The cumulative count of movement in three directions is represented by the percentage occurrences with respect to the total number of pixel movement and stored as a three component vector with the three components representing the distance covered in east, south and west directions, respectively. Similarly right, top and bottom profiles are calculated. Therefore a total of 12 profiles features have been obtained (3 for each profile).

2) Transition Features: In transition features, location and number of transitions from background to foreground pixels in the vertical and horizontal directions are calculated. To get this information, sample image is first scaled to 50*50 and then scanned from right-to-left, left-to-right, top-to-bottom and bottom-to-top. A transition which is close to the starting side is assigned a high value compared to a transition computed at the ending side. For example if the transitions were being computed from right-to-left, a transition found close to the right would be assigned a high value compared to a transition computed to the left. A maximum of five transitions have been recorded in each direction. If there were fewer transitions than the maximum value, then the remaining transitions would be assigned values of 0. It will produce four matrices, two matrices having dimensions W × 5(one for top-to-bottom and other for bottom-to-top) and other two matrices having dimensions H × 5(right-to-left, left-to-right), where W is the width and H is the height of the scaled sample image. After evaluation of transitions each matrix has been divided into five equal parts. We have taken the average of transitions vertically in each part. We got 100 (4 × 25) transition feature vector.

3) Zoning Features: For extracting these features, the sample image has been partitioned into the seven equal size windows both horizontally and vertically called zones. Density value (percentage of black pixels) for each of the zone has been calculated. All these density values have been used to form the input feature set. As we have partitioned the sample image into 49 zones, therefore a density value from each zone makes a feature vector set of size 49.

4) Directional Distance Distribution: For these features for each black/white pixel, nearest white/black pixel is located in eight different directions (0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, 315^0). These distances are then stored in a set of size 16, for each pixel. The first 8 elements of this set correspond to the distance of white neighbors of the black pixel in 8 directions. If the current pixel is white then these elements will be set to 0. Similarly the rest of 8 elements correspond to the black neighbors of the white pixel in 8 directions. If the current pixel is black then these elements will be set to 0. For this feature the image is scaled to the size of 36*36. After obtaining such sets for each pixel, the input image array has been divided into 3 equal parts both horizontally and vertically, hence producing 9 zones. From each zone 16 feature vectors have been obtained by adding the corresponding elements of all the sets, corresponding to the pixels in that particular zone. Therefore 16 features from each zone makes a total of 144 (16*9) feature.

B. Gabor Filter

A Gabor filter is a kind of local narrow band pass filter and selective to both orientation and spatial frequency. It is widely applied in the field of character recognition, face and texture recognition. A two dimensional Gabor filter is defined by the equation (4) given below:

$$f(x, y, \phi, \sigma_x, \sigma_y) = exp\left[-\frac{1}{2} \left\{ \frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2} \right\} \right] \times exp\left\{ i \frac{2\pi R_1}{\lambda} \right\} \quad \ldots (4)$$

Where $R_1 = x cos\phi + y sin\phi$ and $R_2 = -x sin\phi + y cos\phi$

λ and ϕ are the wavelength and orientation of sinusoidal plane wave, respectively. Where σ_x and σ_y are the standard deviations of Gaussian envelop along x-axis and y-axis. In our case $\sigma_x = \sigma_y$. Before feature extraction the image is scaled to the size 32*32. The Gabor feature can be viewed as the response of Gabor filter, which can be obtained by convolving

the filter with an image. A rotation of the x-y plane by an angle ϕ will result in a Gabor filter of orientation ϕ. The value of ϕ is given by $\phi = \pi(k-1)/m,\ k = 1 \ldots \ldots m$, Where m denotes the number of orientations, which are 9 in our case. The filter response corresponding to all orientations are obtained from the whole image, each quadrant and each sub-quadrant of the image, which make a total of 189 features. We have also experimented by increasing the feature vector size to 252, which have been obtained by changing the number of orientations to 12.

C. Discrete Cosine Transform

Discrete Cosine Transform is the member of a family of sinusoidal unitary transforms. Discrete Cosine Transform efficiently encodes energy/the significant details of the image in a few coefficients. These transform coefficients serve as features for the image sample. For the images we have used two-dimensional DCT represented by equation (5). It calculates the two-dimensional cosine transform of an image. In this function M and N are the height and width of the image, but as the image is scaled to the size of 40*40, therefore for this equation M=N.

$$D(i,j) = C(i)C(j)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} p(x,y)\cos\left[\frac{(2x+1)}{2M}i\pi\right]\cos\left[\frac{(2y+1)}{2N}j\pi\right] \quad \ldots.(5)$$

Where

$$C(i) = \begin{cases}\sqrt{\frac{1}{M}} & \text{if } i = 0\\ \sqrt{\frac{2}{M}} & \text{if } i > 0\end{cases} \quad \text{and} \quad C(j) = \begin{cases}\sqrt{\frac{1}{N}} & \text{if } j = 0\\ \sqrt{\frac{2}{N}} & \text{if } j > 0\end{cases}$$

D(i, j) represents the DCT coefficient corresponding to the image pixel p(x, y). Therefore the coefficient corresponding to all the image pixels will constitute a feature vector set. Discrete cosine transform concentrates most of the image energy in very few coefficients. The first transform coefficient is called DC component which is at [0, 0] and rest are called AC components. As the image is scaled to 40*40, therefore a total of 1600 features (transform coefficients) can be obtained from it. But we have picked only 100 features in zigzag manner, as shown in Fig. 3. We have also evaluated the feature-classifier performance by increasing the feature size to 200, merely by selecting the first 200 features in zigzag manner.

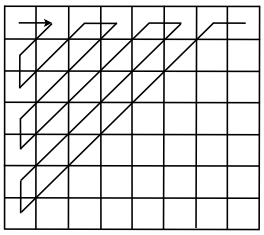


Fig. 3. Zigzag coefficient collection

D. Gradient Features

The gradient features are obtained in three steps: gradient computation, directional decomposition, and feature reduction.

For these features the input image is scaled to the size of 63*63. The gradient vector g(x, y) is then computed at each pixel location using the Sobel operator. Accordingly, the two components; gradient in x and y directions are computed as follow

$$g_x(x,y) = f(x+1,y-1) + 2f(x+1,y) + f(x+1,y+1) - f(x-1,y-1) - 2f(x-1,y) - f(x-1,y+1),$$

$$g_y(x,y) = f(x-1,y+1) + 2f(x,y+1) + f(x+1,y+1) - f(x-1,y-1) - 2f(x,y-1) - f(x+1,y-1)$$

The magnitude and direction of gradient vectors are evaluated from the components g_x and g_y. The gradient vectors are then decomposed into components in eight chain-code directions [8] as shown in Fig. 4(a). If a gradient vector lies between two discrete directions, it is decomposed into two components (Fig. 4(b)) along the two discrete directions; otherwise the magnitude of the vector is exclusively assigned to the corresponding direction. This decomposition results in 63*63*8 values. These values stored in an array are then divided into 81 blocks. The gradient magnitude is accumulated separately in each of 8 directions, for each block, which results in 648 feature values. These values are then down-sampled by using 5*5 Gaussian filter, which reduces the feature vector size to 200.

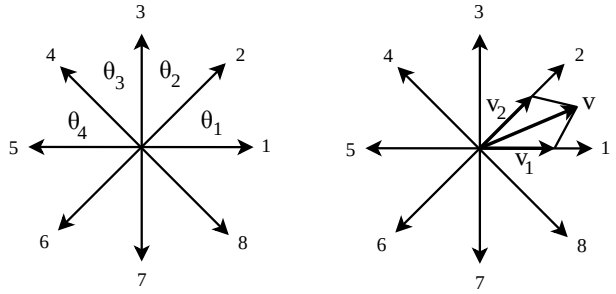


Fig. 4. (a) Eight chain code directions (b) Decomposition of gradient vector

TABLE II. FEATURES AND THEIR VECTOR LENGTH

S.No	Features	Size
1.	Profile direction codes	12
2.	Transition	100
3.	Zoning	49
4.	Directional distance distribution	144
5.	Gabor filter	189/252
6.	Discrete cosine transform	100/200
7.	Gradient features	200

IV. PERFORMANCE EVALUATION

Features of all training and testing samples have been extracted by using above said feature extraction methods. Each

of the classifier explained earlier were then trained by using the training features.

A. Feature-SVM Combination

SVM trained with 7000 training samples has been subjected to classify the 1400 test samples. For classification Linear, Polynomial and Radial Basis Function kernels functions have been employed. We have experimented with these kernels by changing the parameters, like degree of the polynomial kernel, and γ for both polynomial and RBF kernels. Table III shows the percentage classification of the test samples. Here we have shown those values of γ for which the classification correctness is maximum.

TABLE III. PERFORMANCE OF FEATURE(S)-SVM COMBINATION

	SVM			
	Kernels			
Features	**Linear**	**Polynomial**		**RBF**
		***deg* = 2**	***deg* = 3**	
Statistical	**97.714**	97.714 γ=0.003	97.714 γ=0.003	97.143 γ=0.00001
Gabor	97.429	**98.000** γ=0.0005	97.143 γ=0.005	97.143 γ=0.0007
DCT	**98.571**	97.429 γ=0.005	98.000 γ=0.001	96.857 γ=0.003
Gradient	**99.429**	98.857 γ=0.005	98.857 γ=0.005	98.286 γ=0.00005
Stat. + Gabor	97.714	**98.000** γ=0.002	97.714 γ=0.002	96.857 γ=0.00005
Stat. + DCT	97.714	**98.000** γ=0.002	97.714 γ=0.002	97.143 γ=0.00001
Gabor +DCT	**98.286**	97.714 γ=0.003	97.714 γ=0.003	97.429 γ=0.0005
Stat. + Gradient	**99.143**	98.571 γ=0.0019	98.571 γ=0.00001	98.286 γ=0.00005
Gabor + Gradient	**99.429**	98.857 γ=0.0025	98.857 γ=0.0025	98.000 γ=0.00005
DCT + Gradient	**99.429**	98.857 γ=0.0001	97.571 γ=0.00001	98.286 γ=0.00003

B. Feature-ANN Combination

In ANN each output neurons represent the class to be detected. Therefore we have used 70 output neurons. The number of input neurons corresponds to the size of selected feature, and hence number of input neurons can be decided from the size of feature vector length.

The number of the hidden layer neurons was determined experimentally. We started with a number close to mean of input and output neurons and then checked the performance by increasing and decreasing the number of hidden layer neurons. Two functions Elliot and sigmoid have been used as activation function for hidden and output neurons respectively. These function have been determined experimentally form the training data.

TABLE IV. PERFORMANCE OF FEATURE(S)-ANN COMBINATION

Feature	**ANN**		
Statistical	91.143 hidden=100	94.571 hidden=180	**96.571** hidden=360
Gabor	92.000 hidden=90	93.429 hidden=125	**96.286** hidden=230
DCT	95.714 hidden=50	96.571 hidden=85	**97.429** hidden=150
Gradient	91.714 hidden=100	94.000 hidden=135	**96.857** hidden=240
Stat. + Gabor	94.571 hidden=200	94.857 hidden=280	**97.429** hidden=320
Stat. + DCT	92.714 hidden=180	95.714 hidden=230	**96.286** hidden=410
Gabor + DCT	93.143 hidden=80	93.429 hidden=170	**97.714** hidden=330
Stat. + Gradient	94.571 hidden=200	95.429 hidden=287	**97.714** hidden=470
Gabor + Gradient	94.000 hidden=180	93.714 hidden=230	**98.000** hidden=340
DCT + Gradient	91.714 hidden=100	94.286 hidden=185	**97.429** hidden=300

C. Feature-kNN Combination

k nearest neighbor is one of the simplest classification method. Here the Euclidian distance between test-sample feature vector and all of the training-sample feature vectors have been evaluated. And then depending upon the value of k the class of test-sample is predicted. Table V depicts the results of this combination for four different values of k for different feature extraction methods.

TABLE V. PERFORMANCE OF FEATURE(S)-KNN COMBINATION

	kNN			
Feature	***k* = 1**	***k* = 3**	***k* = 5**	***k* = 7**
Statistical	**96.000**	95.429	95.143	94.571
Gabor	**95.714**	95.143	94.571	94.000
DCT	**97.143**	95.143	95.429	96.857
Gradient	**96.571**	96.286	94.857	95.714
Stat. + Gabor	**96.000**	94.857	95.143	95.429
Stat. + DCT	**96.000**	95.429	95.429	95.857
Gabor +DCT	**97.143**	96.571	96.000	96.286
Stat. + Gradient	**96.571**	96.286	95.714	96.571
Gabor + Gradient	96.857	**97.143**	95.714	96.286
DCT + Gradient	**96.857**	96.857	95.714	95.714

D. Effect of Feature Size on Performance

The effects of increase in the feature vector length of Gabor and discrete cosine transform on performance have also been evaluated. In order to increase the feature vector length corresponding to Gabor features for both training and test samples, number of orientations has been increased from 9 to 12. This increases the feature size from 189 to 252. Similarly the numbers DCT feature vectors have been increased by selecting 200 feature vectors from the total of 1600 feature vectors in zigzag manner as shown in the Fig. 3.

TABLE VI. PERFORMANCE OF SVM-GABOR AND SVM-DCT WITH VARYING FEATURE VECTOR LENGTH

	SVM			
	Kernels			
Features	**Linear**	**Polynomial**		**RBF**
		***deg*=2**	***deg*=3**	
Gabor-189	97.429	**98.000** γ=0.0005	97.143 γ=0.005	97.143 γ=0.0007
Gabor-252	97.429	98.000 γ=0.0003	96.857 γ=0.003	97.571 γ=0.0001
DCT-100	**98.571**	97.429 γ=0.005	98.000 γ=0.001	96.857 γ=0.003
DCT-200	97.142	97.714 γ=0.001	97.428 γ=0.005	96.857 γ=0.00005

TABLE VII. PERFORMANCE OF ANN-GABOR AND ANN-DCT WITH VARYING FEATURE VECTOR LENGTH

	ANN		
Gabor-189	92.000 hidden=90	93.429 hidden=125	**96.286** hidden=230
Gabor-252	90.571 hidden=140	92.571 hidden=160	94.857 hidden=290
DCT-100	95.714 hidden=50	96.571 hidden=85	**97.429** hidden=150
DCT-200	95.714 hidden=100	94.429 hidden=135	95.429 hidden=160

TABLE VIII. PERFORMANCE OF KNN-GABOR AND KNN-DCT WITH VARYING FEATURE VECTOR LENGTH

	kNN			
	***k* = 1**	***k* = 3**	***k* = 5**	***k* = 7**
Gabor-189	**95.714**	95.143	94.571	94.000
Gabor-252	95.429	94.857	94.571	93.714
DCT-100	**97.143**	95.143	95.429	96.857
DCT-200	96.857	95.143	94.857	95.714

V. PERFORMANCE COMPARISON

For performance comparison, maximum percentage classification accuracy of all combinations (Tables III, IV and V) has been taken into account.

TABLE IX. COMPARISON OF FEATURE(S)-CLASSIFIER COMBINATIONS

	Classifiers		
Features	**SVM**	**ANN**	**kNN**
Statistical	97.714	96.571	96.000
Gabor	98.000	96.286	95.714
DCT	98.571	97.429	97.143
Gradient	**99.429**	96.857	96.571
Statistical + Gabor	98.000	97.429	96.000
Statistical + DCT	98.000	96.286	96.000
Gabor + DCT	98.286	97.714	97.143
Stat. + Gradient	99.143	97.714	96.571
Gabor + Gradient	99.429	98.000	97.143
DCT + Gradient	99.429	97.429	96.857

Table IX indicates that Gradient feature and its combination with other features, with support vector machines as classifier outperform the others. Discrete cosine transform also perform well with all the classifiers even though it has minimum feature vector length of size 100.

VI. CONCLUSION AND FUTURE SCOPE

From above discussion it has been found that Gradient feature has provided the maximum classification accuracy of 99.429% only with SVM as compared to other combinations. Above results also show that there is no observable increase in the performance with the increase in the feature vector length of Gabor and DCT features.

As the analysis has been done on the isolated recognizable units therefore there may be variation in the results (e.g. due to segmentation process) when these combinations will be used in actual optical character recognition.

The classification results show that different combinations complement each other; therefore as future scope, some methods can be devised to combine the classification outcome of these feature-classifier combinations to improve the classification accuracy of complete recognition system.

REFERENCES

[1] Sinha and H.N. Mahabala, "Machine recognition of Devanagari script," IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-9, pp. 435-441, 1979.

[2] R.M.K Sinha and V. Bansal, "On Devanagari Document Processing," IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1621-1626, 1995.

[3] V.Bansal and R.M.K. Sinha, "Integrating Knowledge Sources in Devanagari Text Recognition System," IEEE Transactions on Systems, Man and Cybernetics-part A: Systems and Humans, vol. 30, No. 4, pp. 500-505, July 2000.

[4] V.Bansal and R.M.K.Sinha, "A complete OCR for printed Hindi text in Devanagari script", Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pp. 800-804,2001.

[5] B.B. Chaudhuri and U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi)," Proceedings of the 4th International Conference on Document Analysis and Recognition, vol. 2, pp. 1011-1015, Germany, 1997.

[6] S. Kompalli, S. Nayak and S. Setlur, "Challenges in OCR of Devanagari Documents," Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), vol. 1, pp. 327-331, 2005.

[7] S. Kompalli, S. Setlur and V. Govindaraju, "Devanagari OCR using a recognition driven segmentation framework and stochastic language models," International Journal on Document Analysis and Recognition. pp. 123–138, 2009.

[8] A. Kawamura, K. Yura, T. Hayama, Y. Hidai, T. Minamikawa, A. Tanaka and S. Masuda, "On-line recognition of freely handwritten Japanese characters using directional features densities," Proceedings of 11th International Conference on Pattern Recognition, Hague, Netherlands, 1992, Vol. II, pp. 183-186.

[9] J. Singh and G.S. Lehal, "Optimizing Character Class Count for Devanagari Character Recognition," International Conference on Information Systems for Indian Languages (ICISIL2011), CCIS vol. 139, pp. 144-149, 2011.

[10] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "*A Practical Guide to Support Vector Classification*," http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. 2010.

Principle of Duality on Prognostics

Mohammad Samie
School of Applied Sciences
Cranfield University
Bedford, UK

Alireza Alghassi
School of Applied Sciences
Cranfield University
Bedford, UK

Amir M. S. Motlagh
Faculty of Science and Technology
University of Westminster
London, UK

Suresh Perinpanayagam
School of Applied Sciences
Cranfield University
Bedford, UK

Epaminondas Kapetanios
Faculty of Science and Technology
University of Westminster
London, UK

***Abstract*—The accurate estimation of the remaining useful life (RUL) of various components and devices used in complex systems, e.g., airplanes remain to be addressed by scientists and engineers. Currently, there area wide range of innovative proposals put forward that intend on solving this problem. Integrated System Health Management (ISHM) has thus far seen some growth in this sector, as a result of the extensive progress shown in demonstrating feasible and viable techniques. The problems related to these techniques were that they often consumed time and were too expensive and resourceful to develop. In this paper we present a radically novel approach for building prognostic models that compensates and improves on the current prognostic models inconsistencies and problems. Broadly speaking, the new approach proposes a state of the art technique that utilizes the physics of a system rather than the physics of a component to develop its prognostic model. A positive aspect of this approach is that the prognostic model can be generalized such that a new system could be developed on the basis and principles of the prognostic model of another system. This paper will mainly explore single switch dc-to-dc converters which will be used as an experiment to exemplify the potential success that can be discovered from the development of a novel prognostic model that can efficiently estimate the remaining useful life of one system based on the prognostics of its dual system.**

Keywords*—Prognostic Model; Integrated System Health Management (ISHM); Degradation; Duality; Cuk Converter

I. INTRODUCTION

Integrated System Health Management (ISHM) [1] is a future advancement incondition based asset management that attempts to create automated prognostic and diagnostic systems to maintain and improve the integrity and readiness expected from legacy Health and Usage Monitoring Systems. ISHM is functioned (over a certain period of time) to detect, predict, diagnose and mitigate adverse events caused by degradation, fatigue and faults in components. For instance, the following problems may occur during an important function related to a system's aircraft, regardless of whether the adverse event had been erupted by one of the subsystems. In order to diligently address this problem, it is important to develop technologies capable of integrating large heterogeneous distributed system [2] and asynchronous data streams from multiple subsystems; hence making it easier to detect a potential adverse event. The following technologies will later be used for diagnosing what caused the event, forecasting what consequences the event will have on the RUL of the system (i.e.,whether the entire system will be put at risk), and lastly to take appropriate precautions to mitigate the event [1].

Moving on, in order to accurately estimate the remaining useful life of devices solely depends on developing prognostic models. This will require additional care and attention to be invested towards preparing the degradation profiles and establishing the physics of failure for every component. Therefore, it's necessary to gather and obtain the degradation profiles of every subsystem, including their individual components. This further result's as a new degradation profile being formulated for whenever a component is upgraded. This degradation profile is calculated from either analysing the accumulated damage or the data driven. A drawback of calculating the degradation profile is that any changes made in the design of the system willconsume time and incur additional costs, since the prognostics model will need to be re-upgraded. It's therefore safe to say that the proposals discussed above are very expensive and consume a lot of time to process while also being unreliable, noisy and inaccurate [3].

We intend to overcome these problems by developing a System – Level Reasoning (SLR) to at least provide the system with significant capabilities that can potentially reduce costs by adding a System Integrated Prognostic Reasoner (SIPR) to the system prognostics [1][4]. For Instance, a Vehicle Integrated Prognostic Reasoner (VIPR) is a project funded by NASA for developing the next generation VLRS. A typical functional module within the SLR is a System Reference Model. The

System Reference Model divides the system into partitions. It also provides the necessary relationships between subsystems, which is required for the inference process. This partitioning allows the inference engine to reuse and link the same prognostic models to multiple subsystems and further minimize certification and qualification costs [1][4].

Although various methods and techniques including: neural network, fuzzy, statistics, semantic computing, graph theory, etc., have been thus far utilized for the development and enhancement of ISHM; however, ISHM continues to suffer from problems related to inefficient models, uncertainties and inadequate reasoning. Additionally, prognostic models also remain to be very costly and time consuming to prepare. The reason these problems exist is mainly because of the systems prognostics heavily relying on the physics of failure models and degradation profiles which are often considered to be inaccurate, inconsistent or even very noisy. We therefore believe that the ISHM system will greatly benefit if the prognostic of a component and system were to be perceived as a feature of a system rather than being perceived as the physics of components. The advantage of this approach is that it will enable SLR to develop prognostics for a new subsystem based on a collection of features (encompassing various models/patterns) already known from the previous prognostics of subsystems; hence saving a lot of time and resources. In order to successfully fulfil this task, SLR may need to employ various techniques associated with Soft Computing (SC),such as fuzzy and neural network within its Inference Engine and System Reference Model units, so that the subsystems properties can be linked to one another. In regards to this project, we expect that a duality connection will be found between the prognostics of dual systems, assuming that the prognostics of dual systems are seen as its parameters and features rather than physics of components.

The prognostics of the system shall be further explained in section 2. The principles of duality in electrical systems, along with brainstorming the duality concept of system's prognostics, are covered in section 3. Section 4 covers the prognostics of dc-to-dc converters with details of Cuk and its dual circuit. The proposed algorithm to develop prognostics for dc-to-dc converters using duality concept is presented in section 5. Simulation results are discussed in section 6. Section 7 discusses future work. Lastly, section 8 concludes the major points discussed in the paper.

II. PROGNOSTICS

In condition-based maintenance, prognostics can be defined as a controlled engineering discipline that focuses on the prediction and estimation of the future course of a system or component that tries to establish when the system/component starts to slowly develop irregularities and faults to the point where it eventually malfunctions. A system or component that malfunctions means that it can no longer operate accordingly. The predicted lifecycle of a system or component is referred to as the Remaining Useful Life (RUL). RUL is used in decision making for contingency mitigation and maintenance. There are various scientific techniques used that help construct the prognostics of a system or component including: failure mode analyses, early detection of aging signs, and damage propagation models.Failure mechanisms are often used in conjunction with system lifecycle management to create prognostics and health management (PHM) disciplines. PHM is also sometimes known as system health management (SHM) or within the field of transportation applications; it is either known as vehicle health management (VHM) or engine health management (EHM). Building prognostic models constitutes of three main technical approaches which fall within the categories of data-driven approaches, model-based approaches, and hybrid approaches [1][4][5].

A. Data-Driven Prognostics

Data-driven prognostics [6] are mainly based on pattern recognition and machine learning approaches that help identify and detect trends and changes in the individual phases of a system's state. A way to predict trends in nonlinear systems is by using classical data-driven methods,such as stochastic models, an autoregressive model, the bilinear model, the projection pursuit, etc. Soft computing techniques that involve using various types of neural networks (NNs) and neural fuzzy (NF) systems have also been commonly adopted to deal with data-driven forecasting of a system state [7][8]. This prognostic approach applies to applications that havecomplicated system architecture, i.e., systems that incur high amount of cost when developing an accurate prognostic model. So by adopting this approach to deal with complex systems will lead the prognostics of a system to be much faster and cheaper to set up as compared to other approaches. Contrarily, data driven approaches may have a wider confidence intervals than other approaches which mean it will require a substantial amount of data for training purposes [9].

There are various strategies used to develop data-driven prognostics which involve the analysis of either (1) modelling cumulative damage and then extrapolating out to a damage threshold, or (2) directly learning from the data based on the remaining useful life.

As it is a lengthy and rather costly process to fail each and every system one by one, we thus seek to obtain the run-to-failure data which refers to the main fundamental setback, especially for new systems. In order to retrieve adequate data-driven prognostics, the accelerated aging data should be extracted cautiously from a number of similar/related products by using appropriate measuring tools. This means that both the quality and quantity aspects of the data driven prognostics will add to expenses; especially since the data sources may have been derived from a wide range of factors including: temperature, pressure, oil debris, currents, voltages, power, vibration and acoustic signal, spectrometric data, as well as calibration and calorimetric data. As a result, it is important to fully understand what parameters and signals will be necessary to measure, and which features will need be extracted from the noisy and high-dimensional data [6][7][9].

B. Model-Based Prognostics

Attempts made towards integrating a physical model of a system which is (either accomplished via micro or macro levels) into the estimated remaining useful life (RUL) is referred to as model-based prognostics [5]. The micro level (also known as material level) is often referred to as damage propagation model which is a physical model that is integrated

in a series of dynamic equations. The following dynamic equations define the very relationships between damage and degradation of a system or component. They further define how the system or component operates under environmental and operational conditions. Despite it being almost impossible to measure many critical damage properties, an alternative solution would be to use sensed system parameters. However, it is possible that the level of uncertainty and inaccuracy to be increased. Even though uncertainty and inaccuracy is added as a result of sensed system parameters, uncertainty management would be considered with the realistic assumptions and simplifications, which may potentially overcome the limitations caused by the sensed system parameter [4][5][10].

In contrast to physical expressions used in micro-levels, macro-level models alternatively use mathematical models at a system level that help define the relationship among system input, system state, and system measurable variables. This mathematical model is often a simplified representation of the system. Simplification may lead to making prototyping faster; but the trade-off to this is that although the coverage of the model is increased, the accuracy of a particular degradation model is consequently decreased. In addition, within a complex application, such as a gas turbine engine, there would be a lack of knowledge in attempting to develop the proper mathematics for all subsystems or components. Again, this leads to uncertainty and inaccuracy, similar to micro-level models; which means simplifications must be considered by performing uncertainty management procedures [1][4][10].

C. Hybrid Approaches

In reality, it is almost impossible to either have a purely data-driven or purely model-based approach. However, both these models do share parts of one another's mechanisms. The intention of hybrid approaches is to show the strength of both 'data-driven' approaches and 'model-based' approaches into one prognostic strategy. Two well-known categories of Hybrid approaches are, 1) Pre-estimate fusion and 2) Post-estimate fusion. The first technique applied, hardly has any 'ground truth' data or 'run-to-failure' data available. The second technique is fitted for situations where uncertainty management is required. This means that the second technique helps to narrow down uncertainty intervals of data-driven or model-based approaches while also improving accuracy [11][12].

III. Prognostics of Dual Systems

Duality is one of the fundamental properties which can be consistently seen in physical systems, such as, electrical, mechanical systems, etc. [13][14]. It has an interesting history in mathematics, engineering and science. Duality relations have been identified between geometric objects, algebraic structures, topological constructs and various other scientific constructs. In regards to electrical systems, duality relations have appeared in the core principles for any theorem within an electrical circuit analysis, for situations where there is a dual theorem that replaces one of the quantities with dual quantities.

Examples of such dual quantities in electrical systems are current and voltage, impedance and admittance, meshes and nodes found in electrical systems (shown in Table 1) [15].

TABLE I. Duality Principle in Electrical Systems

System	Dual of System
Voltage of nodes or across device	Current of branch or mesh
Current of branch or mesh	Voltage of nodes or across device
Resistor (R)	Conductivity (1/R)
Capacitor (C)	Inductance (C)
Inductance (L)	Capacitor (L)
Voltage Source (Vs)	Current Source (Vs)
Current Source (Is)	Voltage Source (Is)
Kirchhoff's Current Law	Kirchhoff's Voltage Law
Kirchhoff's Voltage Law	Kirchhoff's Current Law
Mesh/Loop	Node

In regards to duality concepts, a duality relationship between two electrical circuits is expected to be found, if the values of the parameters and topologies of both circuits are linked to one another based on details in Table 1. Looking at it from a mathematical perspective, dual circuits are known to have the same mathematical model, apart from their parameter differing. Even though the function of systems are different, their prognostics still can be assigned to each other on the basis of dual relationships found between the systems, along with having the same mathematical model with dual parameters shown in Table 1. This provides us with the required facilities to develop the prognostics of a system based on the prognostics of its dual system.

Graph theory [13] well established that the behaviour and the functionality of a system can be recognized by knowing the topology of a system without having to know the components and devicesused in the system (considering we already know the nodes voltages and currents of the branches in the circuits). It can thus be expected that graph theory provides us with the capability to construct the prognostic of a system based on its topology rather than concentrating on the integrated devices and components within the system. It is also to be expected, systems that share similar or dual topologies and mathematical models will also share similar prognostics regardless of the components integrated within the system. This makes it possible to investigate how prognostic models can be constructed from knowing the topology of system rather than having to know thephysics of failure of a system. This therefore makes the process of modelling the prognostics of a system much more feasible and realistic, as it saves a substantial amount of resources and time, since you wouldn't have to go through the process of individually testing every system in order to identify its prognostics.

Fig. 1 shows an example of dual circuits. Using Kirchhoff's laws, it can be said that both circuits have the same mathematical model as shown in equation 1 for circuit in Fig. 1-a; and equation 2 for circuit in Fig. 1-b:

$$Va.(1/R_1 + 1/R_2 + 1/R_3) = 0 \quad (1)$$

$$Ia.(R_1 + R_2 + R_3) = 0 \quad (2)$$

If for instance a degradation mechanisms is added, R_2 in circuit of Fig. 1-b is aged towards a short circuit ($R_2 \rightarrow 0$), this is turned as ($1/R_2 \rightarrow \infty$) in circuit of Fig. 1-a.

In reality, this represents the duality principles shown in Table 1 which proves that the resistor is a dual of a conductive; or in regards to this example, it can be known since the short circuit is a dual of an open circuit.

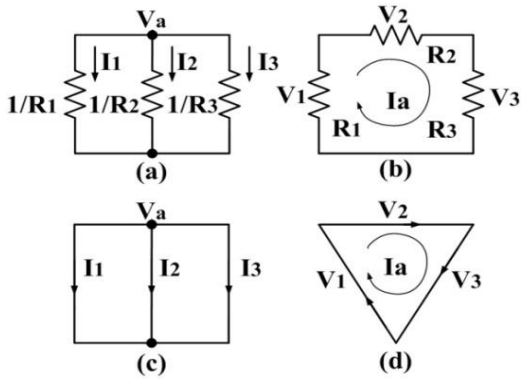


Fig. 1. a, b) Resistive circuit with duality relationship, c,d) Graphs for circuits 1-a and 1-b.

The same rules can be applied to more complicated circuits where various components, including capacitors and inductances are also used. The most critical point that needs to be taken into account is the fact that degradation and failure mechanisms of dual components are not truly related to one another. For instance, the degradation mechanism of capacitor is not in any shape or form related to degradation mechanisms of inductance.

A way to deal with this problem is to rely on the well-known physics principles, such as Ohm's and Kirchhoff's laws. In reference to these two laws, it can confidently be said that any electric component can be formulated by using voltage across the component and current through the component. Alternatively, in regards to graph theory's basic principles of circuit and system design, it has been well known that the behaviour of a system is fully formulated if the voltage of all nodes and current through all the branches in the circuit are also known. This means that behaviour and the function of circuit can be fully formulated no matter what components are used in the circuit, as long as all the voltages and currents are known Fig. 1-c and 1-d, respectively show the graph of the equivalent circuits in Fig. 1-a and 1-b.

Perceiving it from a circuit level, the details required for the development of a prognostics model for a circuit does not necessarily need to be known. Essentially, sensors are used to measure voltages, currents, temperature, etc. By basing it on the meaning of the sensed values, allows the experiences of a degraded circuit or system of any form, to be interpreted as a circuit not functioning properly. Although this principle can be applied for greater purposes, i.e., to design a device independent prognostic model, this paper will for now mainly concentrate on presenting a realization of duality principles for the development of prognostics for dual circuits.

In addition, duality concept has already been recommended for diagnosing faults. Reference [16] proposes a fault diagnoser based on the duality principle and the optimal control theory for linear systems. However, this paper will present duality applications in system prognostics.

IV. Prognostics of DC-To-DC Converter

A basic building block for power convertor type systems is dc-to-dc converter. There are many dc-to-dc voltage and current converters that have various topologies. These topologies can be defined algebraically [17]-[20], graphically, [21][22] or in a matrix form [23]-[25]. It is of significant interest in unifying converter topological characteristics, relationships, and analysis [26]. In regards to health management, the aim would be to develop a basic structure, model or concept that shows where all the converters, including their prognostics may have been derived and mapped. This unified model can lead to many advantages in developing conditional based monitoring, and System- Level Reasoning (SLR).

The underlying concepts related to basic converters can be unified on the basis of what has been already presented in [27] with regard to duality principles and in relation with current and voltage-source converters. The authors in [27] used circuit transformations to unify the basic converters, ultimately showing that other converters are derivable transformation topologies of the basic converter.

This section shows how duality concept can be used to develop prognostic models for Cuk converter and its dual circuit. The following simulations were all conducted with ORCAD and MATLAB. Schematic of Cuk converter and its dual circuit are shown in Fig. 2-a and 2-b.

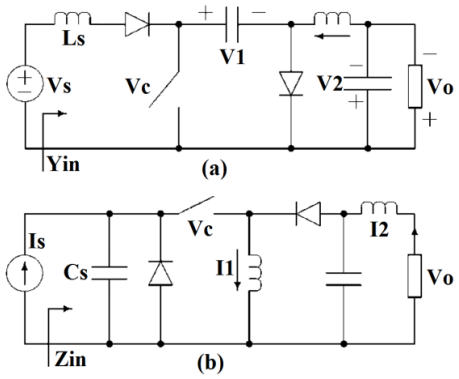


Fig. 2. a) Cuk Converter, b) Dual circuit for Cuk converter in 2-a.

We use certain values for Cuk converter devices, as well as all the equations depicted in reference [28] for all the simulations in this paper. Cuk is a step-down/step-up converter that shares a similar switching topology with buck-boost converter. It thereby presents the voltage ratio of a buck-boost converter [28]:

$$V_o/V_s = D_s/(1 - D_s) \quad (3)$$

where v_o is output voltage, v_g is the input voltage, D_s is the duty cycle of the switch $t_{on}/(t_{on}+t_{off})$; and ton and toff are durations for when the switch is on and off. Equation 3 is calculated from the principle of conservative energy and the fact that the inductor currents relate to the input and output currents. This equation shows that the output voltage can be controlled by maintaining the duty cycle of the switch.

Basedon the type of switching scheme, output voltage can be either higher or lower than the input voltage. The state equations for Cuk converter are:

$$x' = Ax + Bv_g + B_c d \quad (4)$$

$$v_o = C_x$$

$$x = [v_2\ v_1\ i_2\ i_1]'$$

The Cuk converter has two inputs, a control input (V_c) and an input from the power supply (v_s) and one output (v_o). Therefore, matrix [A B C D] relates to 'state space matrices' for the open-loop model from the v_s to the v_o. Similarly, [A B_c C D] is the state space matrices from the control input d to the output v_o. Values for A, B, B_c, C, and D are given in [17]. The same equation can be extracted for dual circuit of Cuk converter in Fig. 2-b; however, parameters are used in a dual form as shown in table 1. Switches in Fig. 2 are IGBT with a control voltage V_c. Y_{in} and Z_{in} are input admittance and input impedance of Cuk circuit and its dual circuit.

Inside converters, the components that are mainly damaged refer to the IGBTs and capacitors. Alghassi et al has discussed different failures mechanisms related to IGBT and they have also presented prognosis model for the dominant failure at a component level in [29][30]. IGBT experiences a number of failure mechanisms including: bond wire fatigue, bond wire lift up, corrosion of the wires, static and dynamic latch up, loose gate control voltage, etc. The resulting affects mentioned are too complex, but we assume that these failure mechanisms can cause IGBT to behave as either an open circuit on a collector-emitter or a device encountering malfunction on its gate-emitter control. For instance, IGBTs thermal junction is increased due to solder crack which turns to wire bond lift off that increases the resistor relating to the collector-emitter. On the other hand, hot carrier injection is increased due to electrical stress. This causes short circuit on the IGBTs gate-emitter junction. The result of this failure, leads to IGBT's gate controllability being missed (loose gate control voltage) causing IGBT to malfunction. The result of this effect is an increase in current through collector-emitter which means that the resistor of collector emitter is decreased. Therefore, it can be established that wire bond lift off and loose gate control voltage are failure mechanisms that presents some kind of duality relationship. While one of them increases the resistor, the other one decreases the resistor. Generally, we assume that IGBT's failure and malfunction mechanisms are parameters that have duality relationships.

Fig. 3 shows IGBT run to failure data relating to four different IGBTs. This data is very noisy and needs to be filtered, but there are still a number of states that can be seen in the data. These states refer to cracks or wires that were lifted up due to degradation mechanisms. The resulting effects are changes in the IGBT's functionality; and changes in the channel resistor of that IGBT. We assume that degradation is processed in a form of duality for Cuk and its dual circuit, so that if IGBT of Cuk experiences degradation towards its open circuit, IGBT of dual circuit of Cuk is degraded towards short circuit.

By the time that the IGBTs are damaged, C_s and L_s are fully charged, as well as the other energy storage components lose energy, so V_o would be 0. It is, however, impossible to have a real short circuit in IGBT, thus we assume that it may have happened when the current through the collector-emitter exceeds over its limit just before the IGBT is burned out.

Based on the level of accuracy, there are number of models that can be applied to a real capacitor and an inductance. To simplify a simulation, we assume that the capacitor and the inductance can both be modelled like Fig. 4 for the purposes of this paper. These models will present duality relationship between capacitance and inductance while also presenting the energy lost by the resistors. R_1 typically has had very large values, while R_2 has a small value; but due to degradation, these resistors are changed towards either open or short circuits.

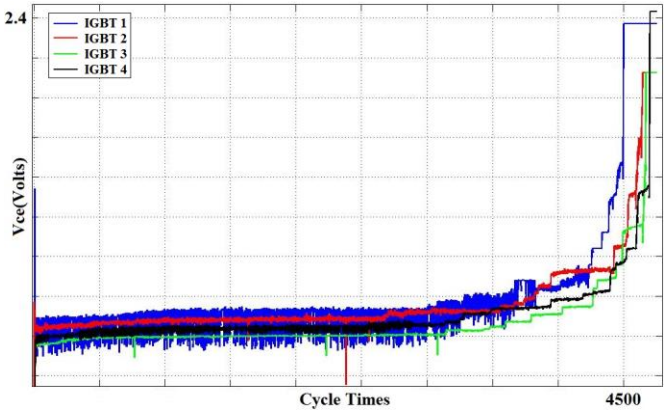


Fig. 3. Real model for a) Capacitor, b) Inductance.

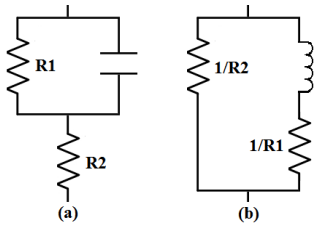


Fig. 4. Real model for a) Capacitor, b) Inductance.

V. Algorithm to Develop Prognostics

Fig. 5 illustrates the proposed algorithm devised to develop this prognostics model. The same process that possesses different sets of run to failure degradation and malfunction profiles is repeated for both Cuk and its dual circuit. At first the components of the circuits are set to be in a good condition. Then as soon as the time step for the circuit is increased, the values of the components are changed by using a series of values provided within the degradation profile for the new time step. Signals, such as v_1, v_2, v_o, i_1, i_2, i_o, are measured at each time step phase.

The following signals are used for calculating the properties of the system, such as transfer functions, input and output impedances and admittances. Subsequently, the system degradation is turned according to changes it has encountered during the transfer functions ($Z_c(d,t)$, $Y_c(d,t)$, $Z_{dc}(d,t)$, $Y_{dc}(d,t)$).

So where d is an index of a selected degradation profile, c is Cuk and dc is the dual circuit of the Cuk converter. Whenever d is altered, time step (t) is reset to zero which resets the process of the circuit to a healthy condition for the new degradation scheme.

The reason for measuring the mentioned signals and parameters is that it would make it possible to understand how energy is transferred between capacitances and inductances; and how that transferred energy is lost when the system is also degraded.

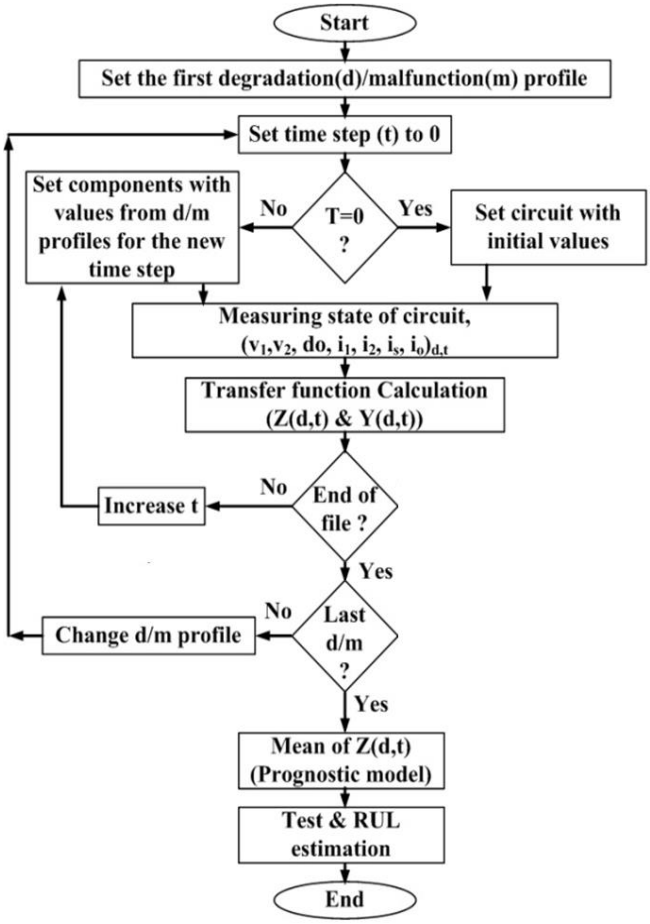


Fig. 5. Algorithm used to develop prognostic model. d/m: degradation/mulfunction number; t: time step or cycle.

To successfully estimate the remaining useful life of both converters, the same process is to be repeated for numbers of different degradation and malfunction profiles. The implication of such a process is to obtain standard Tee and Pi models for Cuk converter and its dual circuit, as shown in Fig. 6. This means that there would be a number of time dependent Tee and Pi models, one for each degradation and malfunction profile. There are many different techniques, such as neural network, fuzzy, statistics, etc., that can be utilised in order to generate a universal prognostic model for the converters (Cuk and its dual) out of all time dependent Tee/Pi models needed to be trained. Here, we just use a mean value to simplify and speed up the process.

The resulted time dependent transfer function which is known as prognostic model is excited with step function ($\alpha u(t)$) during the RUL estimation. Step function $\alpha u(t)$ provides a fixed input of α for the converter over the period of $t>0$. We later assign fuzzy values to the output of transfer function excited with u(t). The fuzzy values represents whether there are a small, medium, normal, transient and big changes experienced at the output of the converter.

The term 'normal' in the fuzzy set, means that changes in the signals can be ignored and transient means that the circuit is in a transient mode and should be settled in a steady state during a specific time constant. RUL is estimated using the MAX fuzzy function which is applied on the triggered fuzzy values. MAX fuzzy function selects the maximum fuzzy value among the fired membership functions.

During this process, RUL is estimated in a fuzzy form, and therefore needs to be de-fuzzified. During the de-fuzzification step, RUL is also scaled up, so that the integration of the estimated RUL (in fuzzy form) reflects the maximum life cycle of the circuit, Fig. 7. Confidence levels are implemented using fuzzy adjectives and adverbs found in fuzzy base knowledge and fuzzy rules.

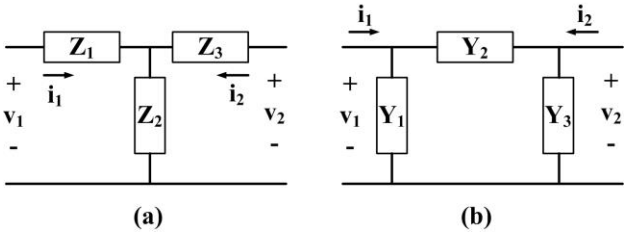


Fig. 6. Tee and Pi Models

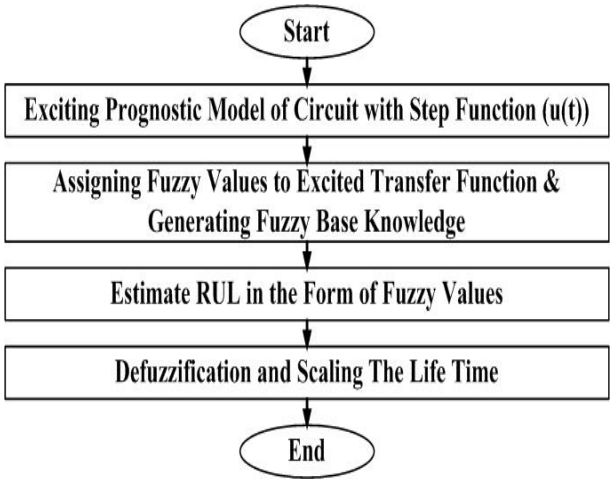


Fig. 7. Simplified Algorithm for RUL Estimation.

VI. Silumation Results and Discussion

Simulation results, using ORCAD 16.6 that reflect the circuits shown in Fig. 3 are presented in Fig. 8 and 9. Looking at these figures, it is apparent that V_{o1} has the same trend as Io2; and the same for Iin1 and V_{in2}.

These in turn reflect the similarities encountered within the transfer functions, such as Z, Y, A_v and A_i shown in Fig. 10. As shown in Fig. 11, we used the IGBT model for our simulations. To add degradation to our simulations, we changed the IGBT's parameters, such as R_{on} in such a way that a trend of failure in Fig. 2 will be obtained from the IGBT model in Fig. 11.

To speed up the simulation, we intend to have all 4500 cycles shown in Fig. 2 in just 25 ms. The result from this mapping is that the degradation will be accelerated in such a way that the first degradation will be experienced around 8 ms after exciting the circuit with step function, u(t); however, the threshold needed to estimate whether the IGBT has aged enough to incur damage in an earlier time is around 10 ms. The same life time and threshold can be expected from energy transfer and power of C_o and L_o, respectively shown in Fig. 12 and 13 for Cuk and its dual circuit.

The figures illustrate that as a result of degradation, energy is not sufficiently transferred in the circuit. So this informs us that the health state of a circuit can just as well be understood from the state equations of a circuit, whereby energy signals from storing elements, capacitors and inductors are used as state variables.

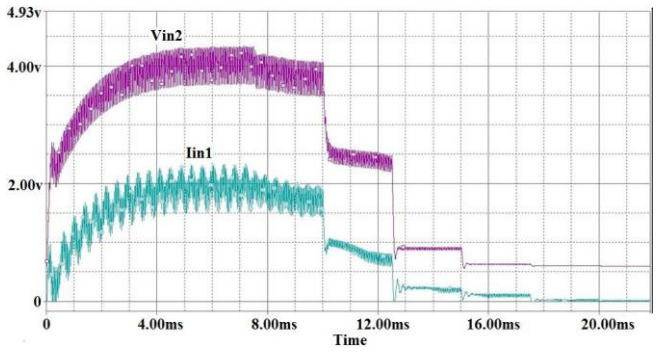


Fig. 8. Changes in Iin1 and Vin2 due to degradation in IGBTs.

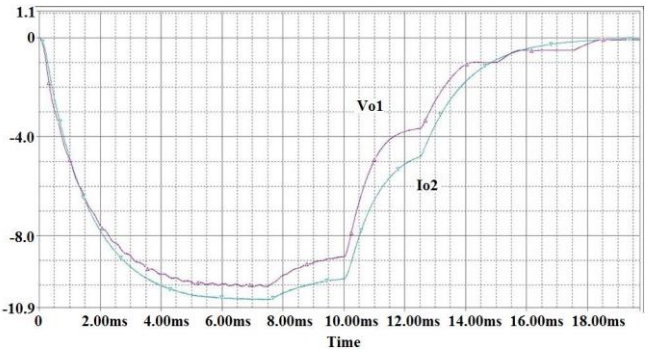


Fig. 9. Changes in Vo1 and Io2 due to degradation in IGBTs.

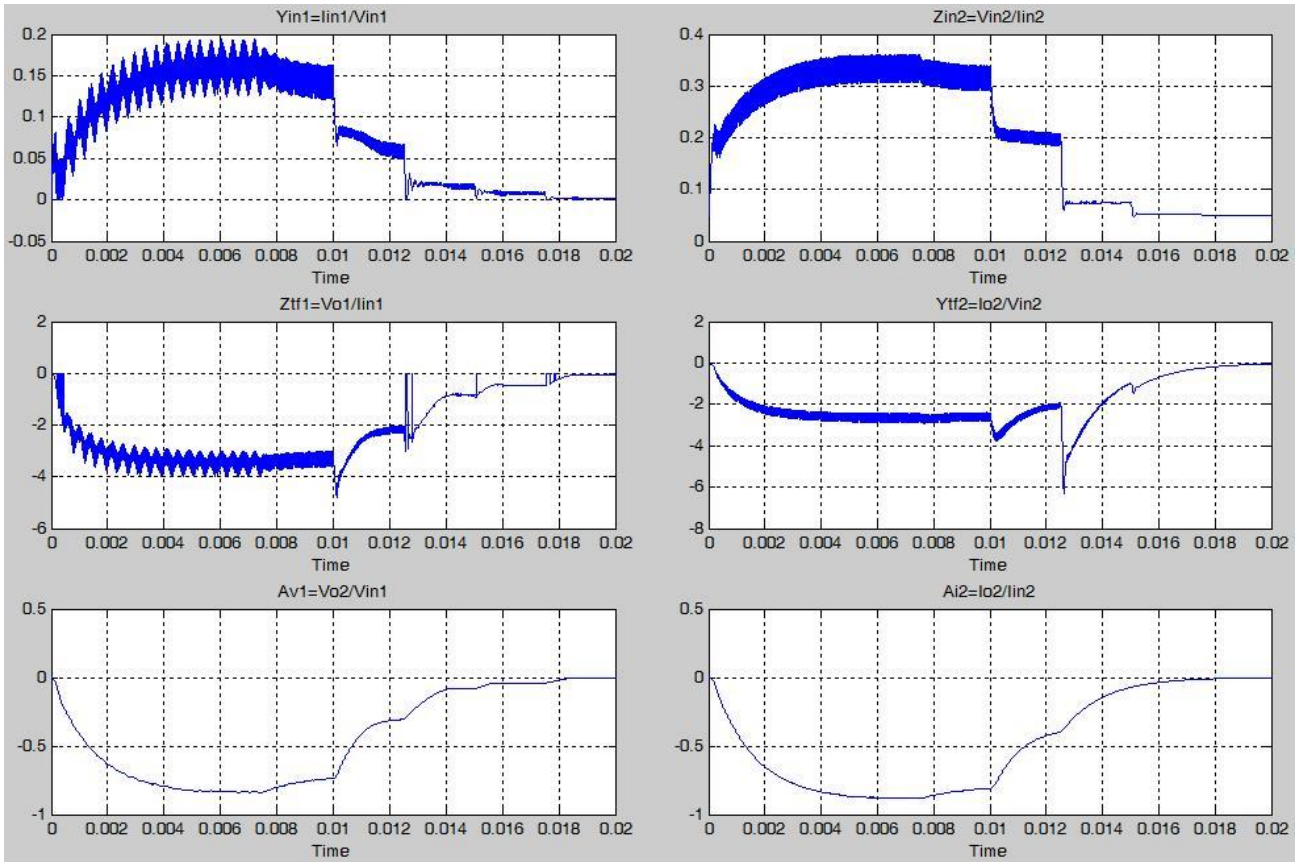


Fig. 10. Changes in transfer functions due to degradation.

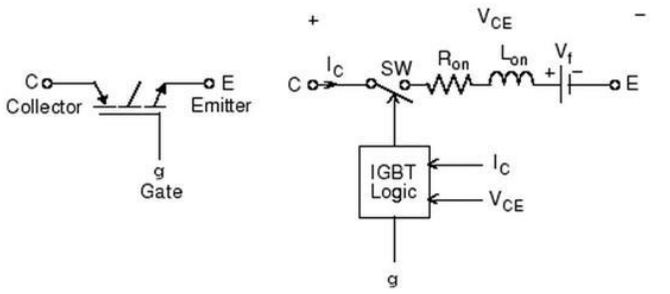


Fig. 11. IGBT Model for Simulation.

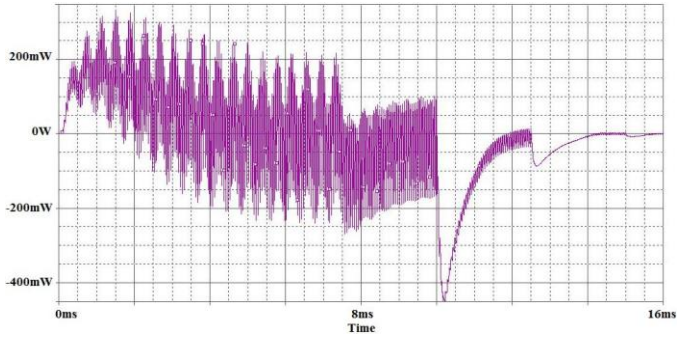


Fig. 12. Energy in C_o of Cuk converter in Fig. 3.

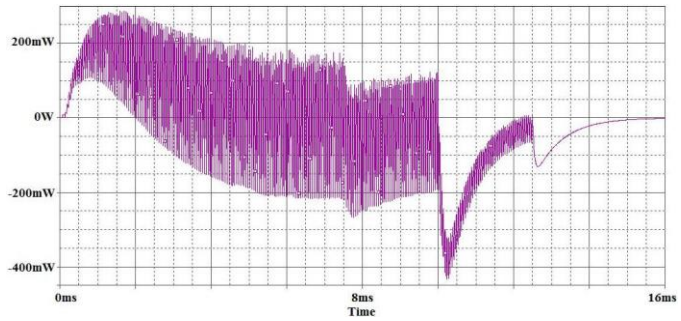


Fig. 13. Energy in L_o for dual circuit of Cuk in Fig. 3.

The timings presented here are not realistic as we accelerated the aging mechanisms in the simulation. Nevertheless, it does clearly prove that the concept of the circuits with duality relationships or even with similarities in components and topologies may also have similar prognostics models.

Every signal, energy and transfer function monitored in the figures have similar information with regard to the effects of degradation of a circuit. However, some of these parameters are rapidly changing due to the switching scheme of IGBT and energy transfer in Capacitors and Inductors, but in most cases the same trend can be found in all these signals. This experiment only refers to the degradation profile concerning the IGBT which refers to the component that mostly experiences degradation during real time; while simultaneously all other components are assumed to behave as non-aged devices (in all simulation). As shown in Fig. 10, A_{v1} and A_{i2} seem to be the best for RUL calculation. Other transfer functions and signals are viewed as noisy data, thus requiring further care to be conducted, such as filtering in order to reduce uncertainties for accurate RUL estimation. For instance, instead of making direct decisions based on monitored signals, the monitored signals can be shifted in the FIFO (First-In First-Out stack) one by one, and the mean value of available data in the FIFO could be hence used for the RUL estimation. FIFO has a fixed storage length, so that shifting a new sample to the FIFO will release the sample that had been already shifted into the stack at the earliest time. Mean value of FIFO captures the trend of signals and eliminates noise, unwanted information and uncertainties. The following FIFO will increase system reasoning within RUL estimation. The length of FIFO has had a great impact on eliminating noise, but it normally shouldn't take that long to lose trend of system degradation. Additionally, implementing a mean value on the stored data in the FIFO may add DC value (i.e., mean value) to the estimated RUL. As DC value is constant (i.e., meaningless information), it will make it easier to eliminate the DC value from RUL.

In order to simplify the process, we use A_{v1} and A_{i2} for the RUL estimation using fuzzy logic techniques to estimate the remaining life time of circuit, as shown from the algorithm in Fig. 7. All the input and output membership functions are set in Gaussian form with input fuzzy values as {small, medium, normal, transient and big} and output fuzzy values as {health_state_1, health_state_2, health_state_3, health_state_4 and health_state_5}. Fuzzy values at input refer to the changes in the trend of A_{v1} and A_{i2}. Fuzzy values at output refer to the life state of circuit, such as young for health_state_1 and, aged for health_state_5. We also use number of adverbs and adjective to address 10% and 90% confidence levels in life estimation. Fig. 14 shows the final RUL in fuzzy form. This figure shows how the circuit behaves in different life/health states during its life cycle. LF_1 to LF_5 show the life domain of each health state. LF_1 comes from having huge transient period at the beginning of A_{v1} and A_{i2}. In reality, we are not faced with such a big transient period, but it is included in our simulation just because of the acceleration in the degradation process. LF_2 represents the how long the system works without the inclusion of degradation, and the rest refer to durations to which the circuit does experience degradation. The reason LF4 appears twice is that the circuit is not experiencing big changes in its transfer functions, but there are meaningful transient periods found that split the LF4 into two separated Gaussian fired fuzzy values.

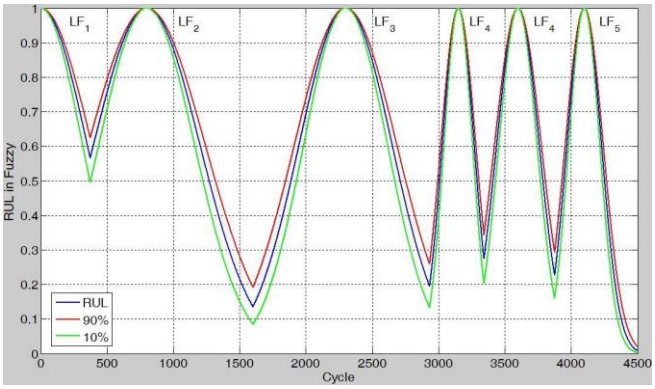


Fig. 14. Resulting RUL after testing prognostic model with data test.

The results reflecting the prognostic model is tested with an additional degradation profile. This will be handled as a test data which will assist us in estimating the remaining useful life time for the converter. Fig. 15 shows the de-fuzzified RUL that represents the remaining useful life with 10% and 90% confidence levels. Ideally, it is expected that the life of a circuit is decreased as a negative ramp in Fig. 15; however, our simulation shows that the RUL is slightly wavy.

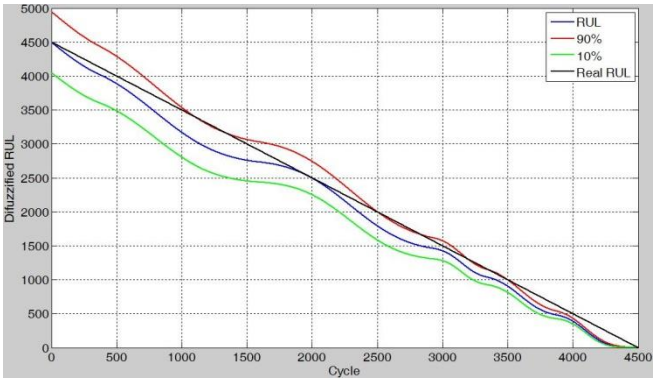


Fig. 15. Resulting RUL after testing prognostic model with data test.

We realized that if a degradation profile is used for Cuk, such that it's converted to a malfunction profile for its dual circuit so that the IGBTs in both circuits always remain in dual forms; then a duality relationship would be found between the transferred functions of these two circuits. For instance, $Z_c(t)$ is equal to $Y_{dc}(t)$. This is because as the degradation profile changes the IGBT of Cuk towards an open circuit; its malfunction profile also changes the IGBT of dual circuit towards a short circuit.

If the malfunction profile for dual circuit of Cuk is not extracted from the degradation profile of a Cuk circuit, then $Z_c(t)$ is not identical to $Y_{dc}(t)$. However, we come to a conclusion that if the whole process is repeated for number of different degradation and malfunction profiles and that the mean value of $Z_c(t)$ and $Y_{dc}(t)$ are used for comparison; leads to meaningful similarity patterns to be found between $Z_c(t)$ and $Y_{dc}(t)$. $Z_{mc}(t)$ can be used for the mean value of $Z_c(d,t)$ and $Y_{mdc}(t)$ can be used for the mean value of $Y_{dc}(d,t)$, in situations where m refers to the mean value. $Z_{mc}(t)$ and $Y_{mdc}(t)$ can be both used as prognostic models for Cuk and its dual circuit. However, these two transfer function are not exactly identical, but they would be more similar to one another if the process that is required to be executed to obtain the functions is repeated for various numbers of degradation and malfunction profiles for both circuits. By implementing more intelligent algorithms that use stochastic, neural network, fuzzy and other techniques instead of a simple mean value function will increase the accuracy of this prognostic model. Implementing such intelligent algorithms also reflects the future aim and direction of our research. Additionally, we should be aware that prognostics have always been a way to estimate the life time of devices and systems within different confidence levels. Confidence levels provide assurance, so that we can comfortably rely on the performance of an aged system. The point is the accuracy of prognostic models has always been under doubt and remains to be under margins of confidence levels. In summary, by using the prognostic model of a system for other systems where similarities in their properties (like duality) are found, would give us a more accurate and reliable representation of the state and condition of the system. This is assuming that the prognostics are developed from adequate number of degradation profiles, and that they also have the right minimum and maximum confidence levels.

VII. Future Work

In this paper, we have looked at the IGBT in a converter as a critical component, thus meaning that the life expectancy of the converter is dependent on the remaining useful life of the IGBT. However, as for another component, such as a capacitor, it is also susceptible to thermal and mechanical stress. Thereby we must investigate whether it is classed as a dominant component failure in a converter or not. So in order to improve the novelty of duality in prognosis, requires one to have a cluster of components. This may overall have a remarkable impact on developing ISHM for critical applications.

VIII. Conclusion

In conclusion, this paper shows that the prognostics of systems that share similar properties in the form of duality can be applied to one another. A prognostic model is developed in the form of a time dependant transfer function where based on the degradation mechanisms related to a system's components, the values are subsequently altered over a certain period of time. So by having the prognostics assigned to a system's property will thereby reflect the duality connection found within the degradation and malfunction profiles of a system. So if we were to consider that the components of a system are aged, this will mean that their dual components in the dual circuit will befaced with malfunction.

The accuracy of the developed prognostic model is highly dependent on the number of degradation profilesavailable; and the methodology used to train the time dependant transfer function. The minimum and maximum confidence levels are used to guarantee and express the accuracy of this model. However, this approach is presented just for Cuk converter and its dual circuit, but it seems that the same technique can be used for systems that have slightly similar mechanisms, properties topologies and degradation. Thereby, further research needs to be conducted for systems that are not in dual forms, especially for the purposes of exploring how the prognostic model of a system could be mapped to the prognostic model of another system.

The advantage and usage of such a technique is emphasized in the implementation stage of the inference engine for System-Level Reasoning (SLR) and System Integrated Prognostic Reasoner (SIPR). It additionally provides us with the required facility to transfer degradation knowledge and experiences between systems. This means that the development of prognostics for huge systems, such as heterogeneous distributed systems used in applications like aircraft will be much faster, while decreasing the cost assigned to accelerated aging tests and preparing degradation profiles. We ultimately intend on pushing forward with our research, in order to apply this technique to the development of the prognostic inference engine and reasoned for aircraft.

ACKNOWLEDGMENT

The authors would like to sincerely thank Professor C Mark Johnson and Dr Paul Evans from the Power Electronics, Machines and Control Group, University of Nottingham for the contribution of failure data of the IGBTs and the power cycling test rig configuration.

REFERENCES

[1] I. K. Jennions, “Integrated Vehicle Health Management Perspectives on an emerging field”, SAE International, Warrendale, pennsylvania, USA 2011, pp. 100-110.

[2] A. El-Sayed and M. El-Helw, “Distributed Component-Based Framework for Unmanned”, Proceeding of the IEEE International Conference on Information and Automation Shenyang, China, June 2012, pp. 45-50.

[3] W. Wenbin and M. Carr, “A Stochastic Filtering Based Data Driven Approach for Residual Life prediction and Condition Based Maintenance Decision Making Support” Prognostics & Systems Health Management, IEEE Conference, Macao, China, Jan. 2010, pp. 1-10.

[4] I. K. Jennions, “Integrated Vehichle Health Management The Technology”, SAE International, Warrendale, pennsylvania, USA, 2013, pp. 139-154.

[5] M. Daigle and K. Goebel, “Model-Based Prognostics under Limited Sensing”, IEEE Aerospace Conference, Big Sky Resort, USA, March 2012, pp. 1-12.

[6] C. Chen and M. Pecht, “Prognostics of Lithium-Ion Batteries Using Model Based and Data-Driven Methods”, 2012 Prognostics & System Health Management Conference, IEEE PHM Conference, Beijing, China, May 2012, pp. 1-6.

[7] H. Chao, D. Byeng, K. Youn, and K. Taejin, "Semi-Supervised Learning with Co-Training for Data-Driven Prognostics", Conference on Prognostics and Health Management, IEEE PHM Conference, 2012, pp. 1-10.

[8] T. Sreenuch, A. Alghassi, S. Perinpanayagam, and Y. Xie; "Probabilistic Monte-Carlo Method for Modelling and Prediction of Electronics Component Life", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014, pp. 96-104.

[9] S. Sarkar, X. Jin, and A. Ray, "Data-Driven Fault Detection in Aircraft Engines With Noisy Sensor Measurements", Journal of Engineering for Gas Turbines and Power, Vol. 133, ASME, August 2011, pp. 1-10.

[10] L. Jianhui, M. Madhavi, K. Pattipati, Q. Liu, M. Kawamooto, and S. Chigusa, "Model-based Prognostic Techniques Applied to Suspention System", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE 2008, Vol. 38, Issue 5, pp. 1156-1168.

[11] P. Shetty, D. Mylaraswamy, and T. Ekambaram, "A hybrid prognostic model formulation system identification and health estimation of auxiliary power units”, Aerospace Conference, IEEE, Big Sky, MT, USA, March 2006, pp. 1-10.

[12] A. K. Garga, K. T. McClintic, R. L. Campbell, Y. Chih-Chung, M. S. Lebold, T. A. Hay, and et al., "Hybrid reasoning for prognostic learning in CBM systems", Digital Object Identifier: IEEE Proceedings on Aerospace Conference, Big Sky, Montana, USA, vol.6, March 2001, pp 2957-2969.

[13] L. O. Chua , C. A. Desoer, and E. S. Kuh, "Linear and Nonlinear Circuits", Mcgraw-Hill, 1st Ed. Edition, March 1987.

[14] D. Y. Gao, "Duality Principles in Nonconvex systems: Theory, Methods and Applications", Kluwer Academic Publishers, USA, ISBN 9781441948250, 2000.

[15] G. E. Sharpe and, G. P. H. Styan, "Circuit Duality and the General Network Inverse", IEEE Transactions on Circuit Theory, Vol 12, IEEE 1965, pp. 22-27.

[16] J. Li, "Optimal Fault Diagnosis Based on Duality Principle for Linear Systems", Control and Decision IEEE Conference, July 2008, pp. 573-577.

[17] R. W. Erickson, “Synthesis of switched-mode converters”,IEEE Power Electron, PESC '83; Annual Power Electronics Specialists Conference, 14th, Albuquerque, NM, June 6-9, 1983, pp. 9–22.

[18] D. Maksimovic, “Synthesis of PWMand quasi-resonant DC-to-DC power converters,” Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 1989.

[19] Y. S. Lee, “A systematic and unified approach to modeling switches in switch-mode power supplies,” IEEE Trans. Ind. Electron., vol. IT-32, no. 4, Nov. 1985, pp. 445–450.

[20] A. K. S. Bhat and F. D. Tan, “A unified approach to characterization of PWM and quasi-PWM switching converters: Topological constraints, classification, and synthesis,” IEEE Trans. Power Electron., vol. 6, no. 4, Oct. 1991, pp. 719–726.

[21] M. Ogata and T. Nishi, “Topological criteria for switched mode dc-dc converters,” in Proc. ISCAS, May 2003, vol. 3, pp. 184–187.

[22] D. H. Wolaver, “Basic constraints from graph theory for DC-to-DC conversion networks,” IEEE Trans. Circuit Theory, vol. CT-19, no. 6, Nov. 1972, pp. 640–650.

[23] Y. Berkovich, A. Shenkman, A. Loinovici, and B. Axelrod, “Algebraic representation of DC-DC converters and symbolic method of their analysis,” in Proc. IEEE 24th Convent. Elect. Electron. Eng., Eilat, Israel, 2006, pp. 47–51.

[24] M. Ogata and T. Nishi, “Topological conditions for passive switches in switching converters,” in Proc. 18th IEEE ECCTD, 2007, pp. 898–901.

[25] T. Nishi, T. Oghishima, and M. Ogata, “Topological conditions on switched mode DC-DC converters,” in Proc. ITC-CSCC, Jul. 2002, pp. 1129–1132.

[26] R. Severns, “Switch mode converter topologies make them work for you,” Intersil, Inc., Milpitas, CA, USA, Appl. Bull. A035, 1980.

[27] B. W. Williams; "Generation and Analysis of Canonical Switching Cell DC-to-DC Converters". IEEE Transactions on Industrial Electronics, VOL. 61, NO. 1, January 2014, pp. 329-346.

[28] F. J. Rytkonen and R. Tymerski, "Modern Control Regulator Design for DC-DC Converters", Electrical and Computer Engineering Department Portland State University. [online]. Available from http://web.cecs.pdx.edu/~tymerski/ece451/Cuk_Control.pdf,2014.05. 14.

[29] A. Alghassi, S. Perinpanayagam, I. K. Jennions; "Prognostic capability evaluation of power electronic modules in transportation electrification and vehicle systems", IEEE, 15th IEEE Conference on Power Electronics and Applications (EPE), 2-6 Sept. 2013, 9 pages.

[30] A. Alghassi, S. Perinpanayagam, I. K. Jennions; "A simple state-based prognostic model for predicting remaining useful life of IGBT power module", IEEE, 15th European Conference on Power Electronics and Applications (EPE), 2013, 7 pages.

Domain Based Prefetching in Web Usage Mining

Dr. M. Thangaraj[1]
Associate Professor, Dept. of Computer Science
Madurai Kamaraj University
Madurai, India

Mrs. V. T. Meenatchi[2]
Lecturer, Dept. of Computer Applications
Thiagarajar College
Madurai, India

***Abstract*—In the current web scenario, the Internet users expect the web to be more friendly and meaningful with reduced network traffic. Every end user needs the channel with high bandwidth. In order to reduce the web server load, the access latency and to improve the network bandwidth from heavy network traffic, a model called Domain based Prefetching (DoP) is recommended, which uses the technique of General Access Pattern Tracking. DoP presents the user with several generic Domains with the top visited web requests in each Domain, which are retrieved from the web log file for future web access.**

Keywords—Latency; Domain; Prefetching; bandwidth; Network Traffic; Web Log File

I. INTRODUCTION

With the unprecedented growth of web, the users always perceive access latency. Intensive measures have been attempted to reduce the Latency. Prefetching is one such approach to reduce the average web access latency. Web Prefetching mainly deals with the ability to identify objects to be pre-fetched in advance. Prefetching is a complementary technique to Caching, which prefetches web documents, that tend to be accessed in near future, while the client is processing the previously retrieved web documents. Various studies have proposed mostly on History based Prefetching.

The interesting and useful access patterns can be analysed and discovered only when web usage data of the user is tracked. This can be achieved only through a branch of web mining, called Web Log Mining. An experiment with a Web Log File of an Educational Institution for predicting the future web requests is attempted here.

This study is divided into five sections each of which deals with a specific issue: Section 1 introduces the subject matter while Section 2 examines various issues associated with Prefetching. Section 3 deals with the Architecture and components of DoP while Section 4 presents the Experimental study and Performance Analysis and finally Section 5 records the concluding remarks.

II. RELATED WORK

Despite the rapid technological advancement in achieving high speed, users demand keeps growing for reducing the access latency. Some of the contributions based on Prefetching, focus on semantic locality, while several do not concentrate on content semantics and specific Domain categorization approach.

The following works do not concentrate on content semantics:

In order to improve the performance of client web object retrieval, the current web page's view time was used for acquiring the web objects of the future web pages. Markov Knapsack method as in [1] was used to define web application Centric Prefetching approach, which restricted the hyperlink domain of webpages to the web application. Though the model accurately represents the client's behavior, considering only the view time of the web page, it is not a wholesome approach.

The importance of preprocessing in Web Usage Mining and the format of the Server Log File is depicted in [14]. Learning algorithm called Fuzzy-LZ as in [7] mines the history of user access and identifies patterns of recurring accesses. To make prefetching decision, a prefetching algorithm based on Neural Network called Adaptive Resource Theory (ART) as in [18] uses bottom-up and top-down weights of the cluster-URL connections.

Various evaluations of analysis of Prefetching performance from user's perspective as in [11] is discussed and the author emphasizes the adaptation of prediction algorithm to the environment conditions. Graph based clustering algorithm as in[10] identifies the clusters of correlated web pages based on the users access pattern in order to improve the proxy server's performance. A group of Prefetching algorithms were reviewed as in [4] based on Popularity, Good Fetch, APL characteristics and Lifetime.

Sequential web access pattern mining as in [20] stores frequent sequential web access patterns in a Pattern tree. The web links generated through Pattern tree are used for recommendations, but they do not concentrate on Domain. User sessions are identified as in [3] and the web logs are cleaned. The user session sequence is generated through Maximum Forward Reference method. The study is defective as it does not focus on semantic locality and the user session sequence is not classified based on their web usage.

A web prefetching algorithm as in [9], particularly concentrated on user's perspective, which analyses the perceived latency with traffic increase and concludes that most likely predicted pages reduce latency. An intelligent solution to caching was proposed as in [17] to improve QOS of websites. It analyses the historical navigation of the website in log file using frequent closed item sets.

Web-object prediction model was developed as in [16] to empower the prefetching engine. It is built by mining the frequent paths from past web log data. Page Rank based Prefetching technique for accessing web page clusters as in [19] deals with the link structure of a requested page and determines the most important linked pages and also identifies the pages to be pre-fetched.

User behavior as in [21] is represented by sequence of consecutive web page accesses from proxy server access log. Indexing methods are used to organize the frequent sequences of the log. The introduction of semantics yields better results. The following citations perform prefetching based on semantics:

Reference [5] introduces a technique which predicts future requests based on Semantic preferences of past retrieved documents in a News Agent Prefetching system. The system extracts the document semantics by identifying keywords in their URL anchor texts. The anchor text for a current web page is associated with so many keywords. Need for more space to store a large set of keywords makes this approach disadvantageous. Selective Markov models as in [15] uses semantic information to prune its states in high order. The system uses semantic distance matrix to store all semantic distances among n webpages in the sequential database. A solution based on Sematic Web Mining was defined in [12] for the Website Key Object problem.

A Website core Ontology was represented for Web user interests. The drawback of the system is that the userr interests may change over the time period. Several methods of prefetching is explained in [13]. Basic scheme of Semantic Prefetching system is discussed. The paper [8] discusses how Semantic Web Mining improve the results of Web Mining.

A Semantic link Prefetcher as in [2], uses the current web page's hyperlink set to trace objects to be pre-fetched during the view time of the current webpage.

Reference [6] uses keyword based semantic prefetching in Internet News. It has taken the News domain alone for prediction. The system analyses the keywords found in the anchor tag for making semantic preferences. That system is known as the keyword method, which is taken up for comparative study.

The following section examines the proposed work, which overcomes the above mentioned problems.

III. PROPOSED WORK

Here a new architecture termed as Domain based Prefetching (DoP), as shown in Fig.1 is proposed.

A. DoP Architecture – An Overview

DoP architecture contains the following four main phases:

1) *PREPROCESS PHASE*
2) *CATEGORIZE PHASE*
3) *ONTOMAP PHASE*
4) *PREFETCHING PHASE*

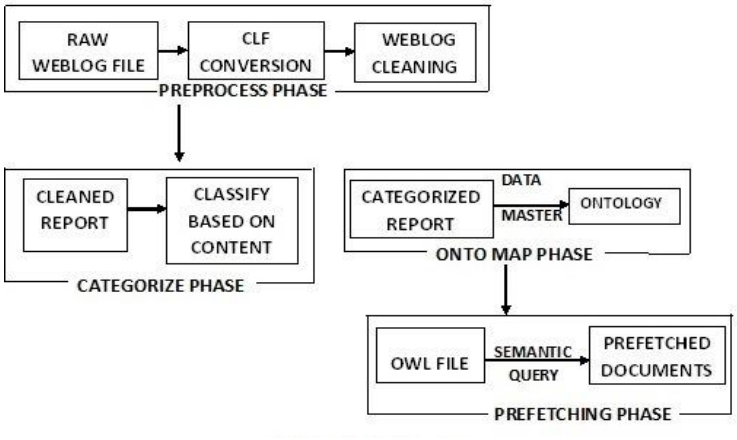


Fig. 1. DoP Architecture

The details of the phases are as follows:

1) *PREPROCESS PHASE:*

This phase concentrates on two main components, DoP Conversion and DoP Cleaning of the web log file.

a) DoP CONVERSION:

The web log file cannot be used as such. With the raw format of the web log file, no useful process can be executed. Of the various web log conversion formats available, the conventional one is the Common Log Format.

Since, the web server's log file does not follow any uniform format for storing entries, it needs to be converted into a format, which would be useful for further processing. This component accepts the raw web log file as its input and converts it into Common Log Format (CLF). Fig. 2 shows the

4030/Apr/201207:14:37pageview/cs497rej/et++/index.html 67.8.221.107 http://www.suksh.com/suk.html - (not authenticated) 200 1 0 0 0 0 0 3.21 k 71/Apr/2012 00:05:21 hit /seized/picts/(nonpage) 4.159.119.132 http://www.flowerfire.com/seized/reviews/blackmantle_sara_lipowitz.html - (not authenticated) 200 0 0 0 0 0 0 .34k 184 30/Apr/2012 09:10:40 page view /cs497rej/et++/index.html 127.8.21.10 - (not authenticated) 200 1 0 0 0 0 0 4.45 k 4 1/Apr/2012 00:04:26 spider /cs497rej/et++/src/(nonpage) 68.142.249.10 - - (not authenticated) 200 0 1 0 0 0 0 1.38 k 40 30/Apr/2012 07:14:37 page view /cs497rej/et++/index.html 17.8.221.107 http://gate.iitd.ac.in/jam.html - (not authenticated) 200 1 0 0 0 0 0 6.12 k 914 30/Apr/2012 17:14:20 page view /cs497rej/et++/index.html 67.8.21.17 http://www.travelsupermarket.com/c/cheap-flights/india.html - (not authenticated) 200 1 0 0 0 0 0 5.11 k 840 30/Apr/2012 12:08:17 page view/cs497rej/et++/index.html 67.8.221.107 http://photobucket.com/pic.html - (not authenticated) 200 1 0 0 0 0 0 3.21 k

sample web log file.

Fig. 2. Sample Web Log File

b) DoP CLEANING:

Web requests include spiders, web robots, files with different extensions other than html, the log entries generated for extremely long user sessions, log entry without proper URL address and requests with status code other than 200, 304, 306 with GET method.

Those web requests need not be considered for further processing and they need to be removed, since they are not useful in mining meaningful knowledge.

2) CATEGORIZE PHASE:

The purpose of this phase is to categorize the web log entries. The Categorization has been done for Semantic Prefetching. From the cleaned web log entries, the URL part is extracted and it is classified based on its content from the corresponding html file through the meta tag. i.e.,<meta name="description"content="">.

Categorization process always needs classifier or class label to perform classification. Here, classifier is the predefined domain name like News, Education, Shopping, Mail. Every entry that is being categorized is placed under the specified domain.

3) ONTOMAP PHASE:

This phase focusses on mapping the classified domains into the Web Ontology file, owl file. This is done through the configured plug in called Data Master of the Protege tool. Ontology will therefore contain the URL and its frequency is termed as HIT under its Domain name.

4) PREFETCHING PHASE:

The ranking of the web request is carried out by taking URL and the hit rate as the sort keys. Ranked web requests under each domain are stored. The prefetch list and purge list are maintained based on the Threshold value, which is based on the value of the hit rate.

All Phases of DoP architecture are interdependent. The basic work flow diagram is presented in Fig. 3, which clearly depicts the placement of Prefetching system in the proxy server.

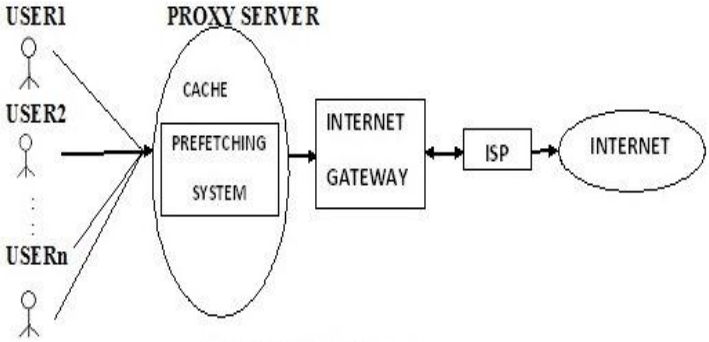


Fig. 3. Work Flow Diagram.

The prefetching system contains the popular web requests, predicted for every domain. User requests which match the predicted requests in the near future might be served from the proxy, without disturbing the original web server, which ensures reduction of the server load and access latency.

The detailed work flow diagram is shown in Fig.4.

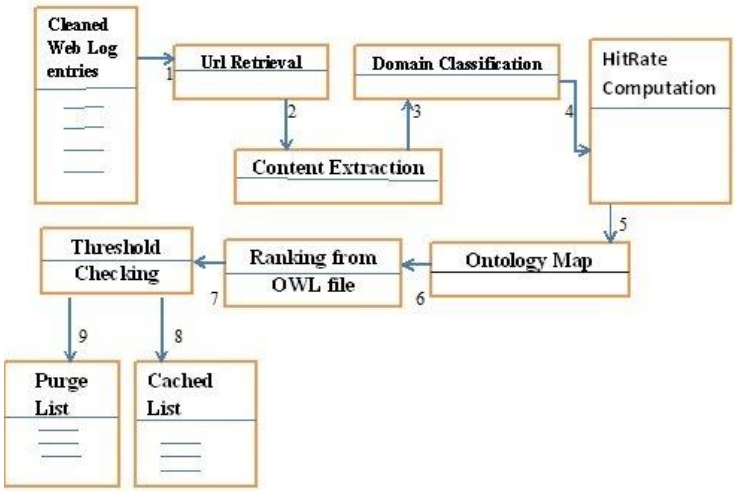


Fig. 4. Detailed Work Flow Diagram

B. Algorithms

The Categorization algorithm- "Categorize" takes the cleaned log file as an input and produces the web ontology file as an output.

Every web request from the cleaned log file is being scanned and the URL segment is tokenized. From the <meta content tag> of the web request, the keywords are fetched and checked with the predefined keyword list. Once there is a keyword match, the web request's URL is stored under its domain, which are then mapped into the Ontology to create 24 distinct classified domains with their corresponding web request's URL which is then mapped into the Ontology.

```
Algorithm      : Categorize(cl,ol)
Data Structure : Table
Input          : cl – Cleaned Log File
Output         : ol – owl File containing classified Domains.
/* i  : 1<=i<=rq  (rq- web entries)
   n  : Total no. of Domains
   url : http request
   Dk  : Array of stored keywords
   K1  : Array of extracted keyword
   hl  : html file of the request
   D   : Domain Table
   Ol  : web ontology file
   kw  : keywords list.*/
for each rq i in cl
{
 tokenize rqi →url;
}
 for each hl of rqi →url
{
  add rqi → kw from <meta name="description" content=" "> to k1[];
for each Dk in n
{
```

```
if (Dk[] = = k1[])
add rqi →url to D;
add D to ol;
}
}
```

The Prefetch algorithm takes web ontology file as its input and produces prefetch cache and purge list as the output. The classified domain contains large numbers of related web requests, of same type. For those entries, the frequency count is computed and stored as the hit rate. The web object's i.e., the url with the corresponding hit rate is then ranked based on the hit rate as the sort key. The sorted web requests are stored under its corresponding domain.

```
Algorithm: Prefetch(ol,dc,pl)
Input    : ol - Web Ontology file with Domains
Output   : dc - cached requests;   pl - purge list
/*url    : http request
min_th   : minimum threshold value
D        : Domain containing classified requests
hr       : hitrate
freq_ct  :function to compute the frequency of web requests
u[]      : Array of http requests
n        : Number of url in D*/
for each url in D
{
  hr  = freq_ct(url);
  sort(url,hr);
  add m to ol;
for each url →hr in ol
{
 if (url →hr <= min_th)
  add url →hr to pl;
else
 add url →hr to dc;
}
}
freq_ct(url)
{
 cn:=0;
for each url in D
{
 if(url == u[])
  cn++;
}
  return;
}
sort(url,hr)
{
m := urli →hr;
for each i in n
{
for each urli in D

{
  if (urli+1→hr > m)
  m:= urli+1 →hr;
}
}
return;
}
```

Threshold value is based on the web object's hit rate. The web object which exceeds the minimum threshold value is stored in the prefetch cache while others are stored in the purge list.

Maintenance of prefetch cache and purge list enables the prefetch cache to contain the most popular web requests and enables the purge list to check periodically with the stored purge list. This is done to permanently remove some web requests, which are consistently retained in the least rank. The purge list is maintained to improve the cache efficiency, since cache can hold only limited web objects.

IV. Experimental Results

DoP approach has been implemented with the use of JAVA, Protégé. The set of experiment explores the web log entries with its various attributes on performance. All experiments were done in Intel Core i5 2.67 GHz with 4 GB RAM, running Windows 7. As an input dataset, the Web Log file of an Educational Institution was analysed. The Log file contained around 1,80,000 entries , collected for a period of 1 year period.

The objectives of the experiments are as follows:

- To improve the proxy server's efficiency. This in turn will reduce the web server load.
- To reduce the user access latency, since the predicted requests are served from the cache, when user request is matched.
- The DoP system suggests the top popular websites in each domain. The web requests under each domain gives clarity to the user when surfing the web.

A. Performance Metrics:

To reduce the access latency, the following four main metrics are vital for prefetching. They are Hit rate, ByteHitRate, Waste Ratio and Byte Waste Ratio.

- HitRate: The percentage of the requested objects serve from prefetching cache.
- ByteHitRate: The percentage of the requested objects serve from the prefetching cache in terms of size.

- WasteRatio : The percentage of undesired documents in the prefetching cache.
- ByteWasteRatio: The percentage of undesired documents in the prefetching cache in terms of size.

The Coverage and Accuracy metrics are also employed.

- Coverage: It is the measure to evaluate the efficiency of prefetcher in satisfying the future object request demand.
- Accuracy: It is the measure of the total prefetched objects, actually used to satisfy the user requests from the prefetched objects.

B. Equations:

Domainwise Coverage is calculated by using the formula given in (1),

$$C_i = C_n / n \quad (1)$$

where,

C_i is the coverage metric.

C_n is the total number of objects in the specific domain d_i.

n is the total number of web objects in cleaned log file.

Domain wise Accuracy is computed as given in (2),

$$A_i = u(d_i) / C_n \quad (2)$$

where, $u(d_i)$ is the total number of objects used in each domain.

The Hit rate percentage(*hr)* is computed as given in (3),

$$hr = 100 * A_i \quad (3)$$

where, A_i is the Accuracy.

Parameters taken for our study is given in Table 1.

TABLE I. PARAMETERS USED IN DOP SYSTEM.

Parameter Name	Description
Wl	Web log file ranges from 1 to 1,80,000 entries.
d_n	Domain Name is of string value (News, Education, Advertisement)
N	Total no. of domains, for the study is 24, which may be increased
Ul	http request of the log file.
Pc	prefetch cache, file that stores the popular web requests
Pl	purge list, file that stores the web requests, to be removed after threshold consideration.
min_th	Minimum threshold value based on hit rate.
max_th	Maximum threshold value based on hit rate.

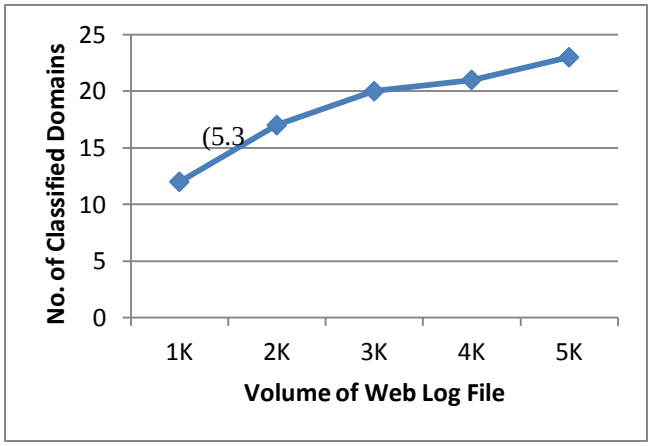
Fig. 5 presents the data size of the web log file with the variation in Throughput. Throughput is the time measured in millisecond, which includes the total time taken for the Log file cleaning, CLF conversion, Log entries categorization, Ontology mapping and Prefetching.

Fig. 5. Throughput Analysis

(5.1)

Categorization efficiency is achieved only when the log entries are correctly classified under its domain.

Fig.6 shows the no. of classified domains with the corresponding log entries. This study has 24 fixed domains for Categorization.



(5.3

Fig. 6. Categorization Analysis

Fig. 7 clearly shows the distribution of the web requests hit rate. This is processed from the whole log file. Since it is a Educational Institution Log file, major distribution is towards Education category and Job Search.

From this visualisation, one could easily find the top most popular domain and the least used one. 10%of Others category shows the ratio of the unclassified web requests with the total web requests. The reduced percentage in Others category reveals the categorization effectiveness.

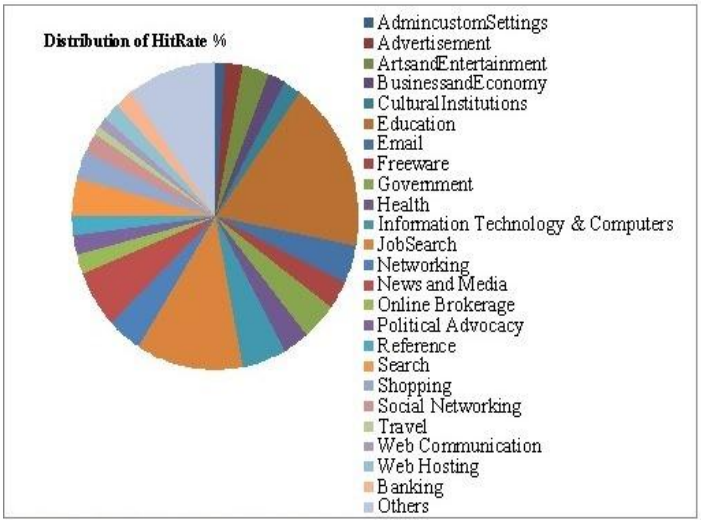


Fig. 7. Overall hitrate distribution for 24 domains

B. Comparitive Study:

In this section, DoP method is compared with the KW method, which considers the News domain. To be generic, the proposed system takes 24 domains into account. Major 4 metrics of prefetching were considered for comparing the DoP method with Keyword based method. Fig. 8 shows the comparative study of DoP with Keyword based method in terms of Hit ratio.

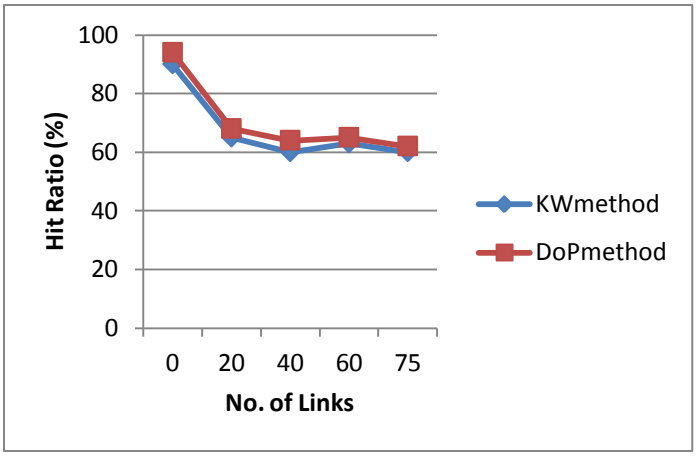


Fig. 8. Growth of Hit Rate against No.of Links.

From Fig. 8, considerable improvement in hit rate of DoP method is clearly learnt. Fig. 9 shows the comparison of DoP with Keyword based method in terms of byte hit rate. The byte hit rate is based on individual web object size. The increased percentage in byte hit rate of DoP method, shows that large number of objects have been requested and fetched from the web log file.

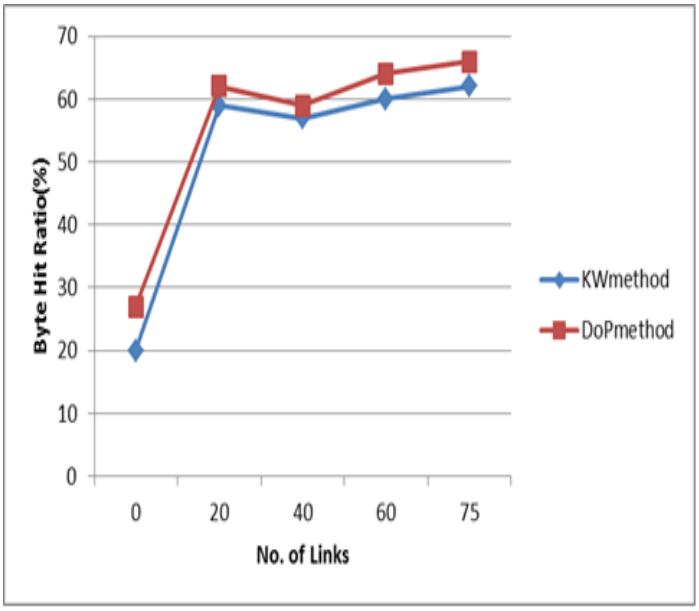


Fig. 9. Performance on ByteHinRate

In DoP method, the no. of undesired document s in theprefetch cache is computed with the help of the purge list. Fig. 10 shows the Waste Ratio comparison.

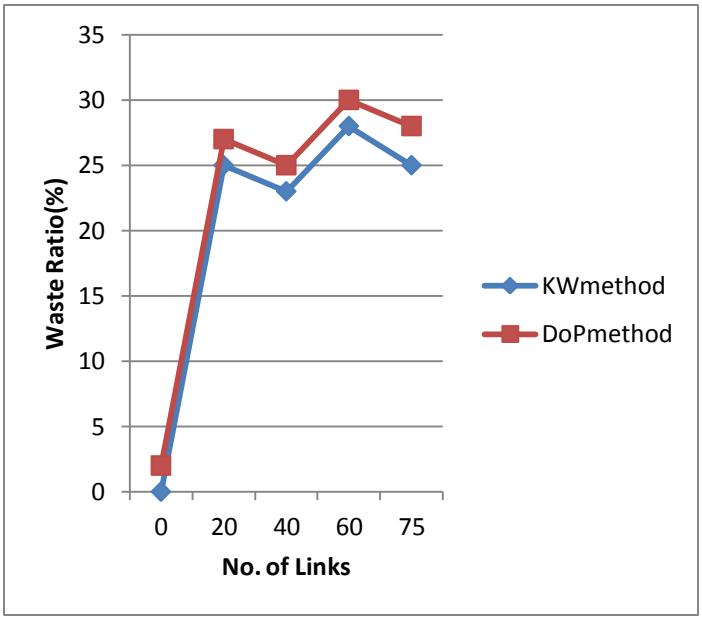


Fig. 10. Waste Ratio Analysis

The associated size of the undesired web objects that reside in the prefetch cache is the Byte Waste Ratio as shown in Fig.11.

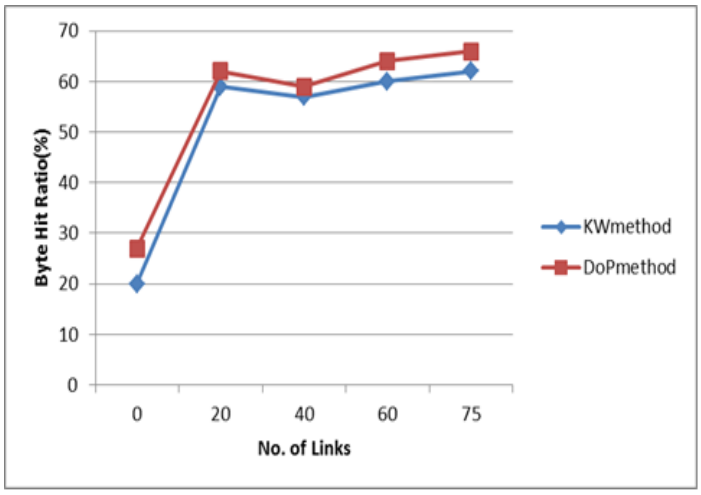


Fig. 11. Byte Waste Ratio Analysis

V. CONCLUSION

This study has presented architecture of Domain based Prefetching in Semantic Web and gives importance to the need for Content Prefetching and Domain wise Prefetching. The system facilitates the user request from relevant cluster. The performance aspect shows the DoP method which outperforms the existing method with varied domains and achieves the hit rate of 80%. The system considerably reduces the access latency. Since the log file of an educational institution was taken as the main platform, the set of users and their interests do not vary in different sectors. The user access patterns are almost decisive in nature. Currently this web ontology file is used for mapping of web log entries, SPARQL queries were used for retrieval and prediction.

This Research may further be extended by periodical analysis of the other domain. Instead of owl file representation, RDF structures may be used for representing the log file and the individual entries of the log file may be annotated. The current study focusses mainly the generic domains, further if an individual domain is separately analysed, there is a lot of scope to prove with more constructive results. There is innumerable number of areas available for further exploration in Prefetching.

REFERENCES

[1] Alexander, P.Pons,"Improving the performance of client web object retrieval," The journal of the Systems and Software 74, 2004 pp. 303-311, doi: 10.1016/j.jss.2004.02.030.

[2] Alexander, P. Pons, "Object Prefetching Using Semantic Links," The DATABASE for Advance in Information Systems," 2006 Vol.37, No. 1.

[3] Arumugam, G and S.Suguna, "Predictive Prefetching Framework Based on New Preprocessing Algorithms Towards Latency Reduction," Asian journal of Information Technology 7(3) , 2008 pp. 87-99, issn: 1682-3915.

[4] Bin Wu and Ajay D. Kshemkalyani, "Objective - Optimal Algorithms for Long-Term Web Prefetching," Proc. of IEEE Transaction on Computers, 2006 Vol. 55, No. 1.

[5] Cheng-Zhong Xu and Tamer I.Ibrahim, "Semantics-Based Personalized Prefetching to Improve Web Performance," Proc. of the 20th IEEE Conf. on Distributed Computing Systems, 2000 pp. 636-643.

[6] Cheng-Zhong Xu **and** Tamer I.Ibrahim, "A keyword-based semantic prefetching approach in Internet news services," Proc. of IEEE Transaction on Knowledge and Data Engineering, 2004 doi: 10.1109/TKDE.2004.1277820

[7] Daby M. Sow, David P. Olshefski, Mandis Beigi and Gurudth Banavar, "Prefetching Based on Web Usage Mining," International Federation for Information Processing 2003, LNCS 2672, pp.262-281.

[8] Gerd Stumme, Andreas Hotho and Bettina Berndt, "Semantic Web Mining State of the art and future directions," Elsevier Journal of Web Semantics 2006 doi: 10.1016/j.websem.2006.02.001.

[9] George Pallis, Athena vakali and Jaroslav pokorny, "A Clustering - based Prefetching scheme on a Web cache environment," Computers andElectricalEngineering2008pp.309-323, doi:10.1016/j.compeleceng.2007.04.002.

[10] Joseph Domenech, Ana Pont, Jose A. Gil and Julio Sahuquillo, "Guidelines for Evaluating and Adapting Web Prefetching Techniques,".XVII Jornadas De Paralelismo – Albacete 2006, Spain.

[11] Joseph Domenech, J.A. Pont, J. Sahuquillo and J.A Gil, "A User-focused evaluation of web prefetching algorithms,"Computer Communications2007pp.2213-224,doi:10.1016/j.comcom.2007.05.003.

[12] Juan D. Velasquez, Luis E. Dujovne and Gaston L'Huillier," Extracing significant Website Key Objects: A Semantic Web Mining Approach," Elsevier Journal of Engineering Applications of Artificial Intelligence 2011, doi: 10.1016/j.engappai.2011.02.001.

[13] Lenka Hapalova and Ivan jelinek, "Semantic web access prediction," Proc. Of International Conference on Computer Systems and Technologies 2007, ISBN: 978-954-9641-50-9.

[14] Marathe Dagadu Mitharam, "Preprocessing in Web Usage mining," International Journal of Scientific & Engineering Research, February 2012 Vol. 3, No.2.

[15] Nizar, R. Mabroukeh and C.I. Ezeife, " Semantic-rich Markov Models for Web Prefetching," Proc. of IEEE International Conference on Data Mining Workshops 2009, doi: 10.1109/ICDMW.2009.18.

[16] Qiang Yang and Henry Hanning Zhang, "Integrating Web Prefetching and Caching Using Prediction Models," World Wide Web 2001 pp. 299-321

[17] Samia Saidi and Yahya Slimani, "Enhancing Web Caching Using Web Usage Mining Techniques," Springer- Verlag Berlin Heidelberg 2010 pp. 425-435.

[18] Toufiq Hossain Kazi, Wenying Feng and Gongzhu Hu, "Web Object Prefetching: Approaches and a New Algorithm," Proc. Of IEEE International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2010, doi:10.1109/SNPD.2010.28.

[19] Victor Safronov and Manish Parashar, "Optimizing Web Servers Using PageRank Prefetching for Clustered Accesses," World Wide Web: Internet and Web Information Systems 2002 pp. 5, 25-40.

[20] WANG Xiao-Gang and LI Yue, "Web Mining Based on User Access Patterns for Web Personalization," Proc. of ISECS International Colloquium on Computing, Communication, Control and Management 2009.

[21] Yi-Hung Wu and Arbee L. P. Chen, "Prediction of Web Page Accesses by Proxy Server Log," World Wide Web 2002 Vol. 5 No. 1, pp. 67-88.

Teaching Introductory Programming

Agent-based Approach with Pedagogical Patterns for Learning by Mistake

Ljubomir Jerinic
Department of Mathematics and Informatics
Faculty of Science, University of Novi Sad
Novi Sad, Serbia

***Abstract*—From the educational point of view, learning by mistake could be influential teaching method, especially for teaching/learning Computer Science (CS), and/or Information Technologies (IT). As learning programming is very difficult and hard task, perhaps even more difficult and extremely demanding job to teach novices how to make correct computers programs. The concept of design pedagogical patterns has received surprisingly little attention so far from the researchers in the field of pedagogy/didactics of Computer Science. Design pedagogical patterns are descriptions of successful solutions of common problems that occur in teaching/learning CS and IT. Good pedagogical patterns could help teachers when they have to design new course, lessons, topics, examples, and assignments, in a particular context. Pedagogical patterns captured the best practice in a teaching/learning CS and/or IT. They could be very helpful to the teachers in preparing their own lessons. In this paper a brief description of special class design of pedagogical patterns, the group of patterns for learning by mistakes, is presented. In addition, usage of *helpful* and *misleading* pedagogical agents, which have been developed in Agent-based E-learning System (AE-lS), based on pedagogical pattern for explanation *Explain*, and pedagogical pattern for learning by mistakes *Wolf, Wolf, Mistake*, is described.**

Keywords*—*Pedagogical Pattern; Pattern Design; Learning; Programming; Computer science education; Programming; Software agents; Electronic learning; Computer aided instruction

I. INTRODUCTION

Conventional pedagogy believes that the one of good way to teach students is to have them repeatedly practice some tasks. In recent work of Lindsey E. Richland, Nate Kornell and Liche Sean Kao [1] the advantages of learning through error was discussed. According to this approach, it is important to avoid mistakes while learning so that our mistakes are accidentally reinforced. That approach assumes that the best way to teach children is to have them repeatedly practice (test for example) as far as it takes.

Once they know (learn or guess or rich somehow) the right answer, that correct response is embedded into the brain. However, this error-free process turns out to be inefficient: Students learn material much faster when they made mistake first, especially in programming. In other words, getting the wrong answer helps us remember the right one.

Nobody likes making mistakes. Nevertheless, unless you want to go through life as a complete recluse, you are guaranteed to make one every now and them. If you learn from mistakes correctly, they could push you forward. You can only learn from a mistake after you admit you have made it, or get the explanation way you have made it.

However, from the educational point of view, learning by mistake could be powerful teaching technique and/or method. If the lecturer[1] create appropriate situation and put student in it, where student can make interesting mistakes, it could be used for educational purpose, and this method is called the learning by mistake technique of teaching. Of course, the lecturer could use some fine facts to make students to made mistake, and after explanation way you made it, you learn, i.e. do not make the same error again.

Joseph Bergin [2] defined pedagogical patterns as follows "Patterns are designed to capture best practice in a specific domain. Pedagogical patterns try to capture expert knowledge of the practice of teaching and learning. The intent is to capture the essence of the practice in a compact form that can be easily communicated to those who need the knowledge. Presenting this information in a coherent and accessible form can mean the difference between every new instructor needing to relearn what is known by senior faculty and easy transference of knowledge of teaching within the community."

This paper covers one point of view in design and implementation of Pedagogical Patterns, the group of patterns for learning by mistakes method in teaching.

The rest of this paper is organized as follows. Section 2 provides an overview of the existing theory and application related to teaching/learning by mistakes. In the field of e-learning and tutoring systems, two categories of software agents are of the special interest: harvester and pedagogical agents. Section 3 provides an overview of the existing work related to e-learning systems and pedagogical agents.

Section 4 introduces pedagogical patterns, pattern language for describing patterns, and pedagogical pattern Explain, and two distinct sub-types of pedagogical agents: helpful and misleading is introduced. Whereas helpful agents provide the correct guidance for a given problem, misleading agents try to steer the learning process in a wrong direction, by offering false hints and inadequate solutions. The rationale behind this approach is to motivate students not to trust the agent's instructions blindly, but instead to employ critical thinking, and, in the end, they themselves decide on the correct solution to the problem in question.

[1] In this paper term lecturer is used to denote teachers, professors, instructors, tutors, i.e. it denotes the person who teach.

In Section 5, a stand-alone e-learning architecture, called Agent-based E-learning System (AE-lS) and some examples are described. AE-lS are designed to help learners in learning programming and programming languages. In Section 7, describe design and definition of pedagogical pattern for learning by mistakes Wolf, Wolf, Mistake. Some examples of use that pedagogical pattern is presented in Section 8. Finally, overall conclusions and future research directions are given in Section 9.

II. Teaching/Learnig by Mistakes

For years, many educators have championed "errorless learning," advising teachers (and students) to create study conditions that do not permit errors. For example, a classroom teacher might drill students repeatedly on the same multiplication problem, with very little delay between the first and second presentations of the problem, ensuring that the student gets the answer correct each time.

People remember things better, longer, if they are given very challenging tests on the material, tests at which they are bound to fail. If students make an unsuccessful attempt to retrieve information before receiving an answer, they remember the information better than in a control condition in which they simply study the information [1]. Trying and failing to retrieve the answer is actually helpful to learning. It is an idea that has obvious applications for education, but could be useful for anyone who is trying to learn new material of any kind.

Lecturer could ask students (students could try to answer) questions at the back of the textbook chapter, or to give them eLearning topic test, before teaching and students could try to answer. If there are no questions available, lecturer could convert the section headings to questions. For example, if the heading is Loop-Control, ask students "What is Loop-Control?" If the answers are wrong, teach the chapter/topic and ask the same questions, when the lecture is finished. If the answers are good lecturer should praise students. If the answers are wrong, lecturer gives instructions, extra questions, hints, and discuss why the answers are wrong. For answers that are very wrong, lecturer gives students additional time to try to learn and master the material lectured. Even if answers are wrong, these mistakes are more useful to the students, much more valuable than just learning the material. Getting the answer wrong is a great way to learn.

These are general-purpose strategies for teaching/learning by mistakes, and it is used for design of pedagogical pattern *Wolf, Wolf, Mistake*, described in Section 6. Moreover, this strategy is employed and utilized for *helpful* and *misleading* pedagogical agents, described in Section 5.

III. Teaching Programming With Patterns and Agents

Software agents, or simply *agents*, can be defined as *autonomous* software entities, with various degrees of *intelligence*, capable of exhibiting both *reactive* and *pro-active* behavior in order to satisfy their design goals. From the point of e-learning and tutoring systems, two types of agents are of the special research interest: *harvester* and *pedagogical* agents. Harvester agents are in charge of collecting learning material from online, often heterogeneous repositories [3].

Haake and Gulz [4] define pedagogical agents as "lifelike characters presented on a computer screen that guide users through multimedia learning environments" (p. 28). Heller and Procter [5] points out that main goal of usage of pedagogical agents are to motivate and guide students through the learning process, by asking questions and proposing solutions.

A stand-alone e-learning architecture, called ***A**gent-based **E-l**earning **S**ystem* (***AE-lS***). ***AE-lS*** are designed to help learners in learning programming and programming languages. ***AE-lS*** consist of three main components:

- Harvest*e*r agents;
- Classifier module*;* and
- A pair of pedagogical agents.

The harvester agents are in charge of collecting the appropriate learning material from the web. Their results are fed into the *Classifier* module, which performs automatic classification of individual learning objects. Finally, a pair of specially designed pedagogical agents - one *helpful* and one *misleading* - is used to interact with students and help them comprehend the underlying learning material.

The helpful pedagogical agent provides useful hints for the solution of the given problem to the student, trying to direct student to the correct solution, or to help student to understand some topic, giving explanations. On contrary, misleading pedagogical agent try to steer and guide the solving/learning process in the "wrong" direction, giving some hints or explanation which could produce bed results. The student is never sure with which agent (s)he is interacting, this approach encourages students not to follow the agent's/tutor's instructions blindly, but rather to employ critical thinking and, at the end, they themselves decide on the proper solution to the given problem or the suitable accepting and understanding presented topic.

Originally, the ideas of using harvester, as well as the two types of pedagogical agents were discussed in [6]. This paper presents a concrete implementation of these ideas, in connection with pedagogical pattern approach.

IV. Pedagogical Patterns

What are Pedagogical Patterns? Patterns are designed to capture best practice in a specific domain. Pedagogical patterns [2] try to capture expert knowledge of the practice of teaching and learning. The intent is to capture the essence of the practice in a compact form that can be easily communicated to those who need that knowledge and experience. In essence, a pattern solves a problem. This problem should be one that recurs in different contexts. In teaching, we have many problems such as motivating students, choosing and sequencing appropriate materials and resources, evaluating students, and the similar.

These problems do recur and in slightly different form each time. Each time a problem, pops up there are considerations that must be taken into account that influence our choice of solution. These forces push us toward or away from any given

solution to a problem. A pattern is supposed to present a problem and a solution. The problem together with the forces must apply to make that solution beneficial to the problem.

A. Pattern Languages - The Pattern Format

A pattern language is a set of patterns that work together to generate complex behavior and complex artifacts, while each pattern within the language is itself simple. Pattern languages, on the other hand, promise to drive fundamental and lasting improvements. One very successful pedagogical pattern language is Seminars by Astrid Fricke and Markus Vöelter [2]. It describes how to design and deliver a short course. Little in this language (or any pattern language) is novel, but it brings together in one place expert knowledge that is often forgotten and sometimes overlooked.

Besides its title, a pattern contains at least the following five sections:

- The Context section sets the stage where the pattern takes place.
- The Problem section explains what the actual problem is.
- The Forces section describes why the problem is difficult to solve.
- The Solution section explains the solution in detail.
- The Consequences (positive and negative) section demonstrates what happens when you apply the solution.

The Figure 1 shows the pattern sections and the order in which the pattern should be written.

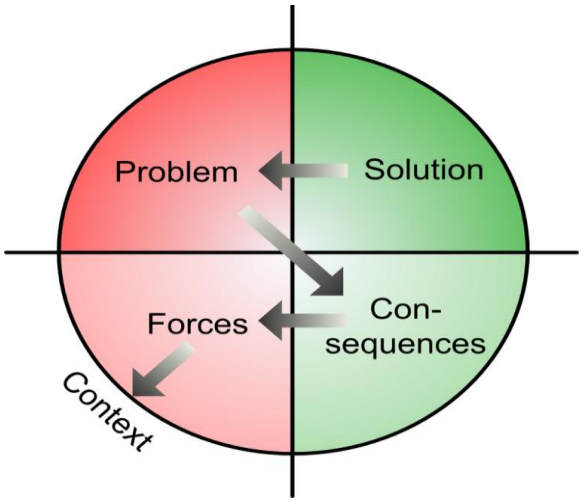


Fig. 1. Pattern language sections and their writing order

B. Explanation Pattern for Explanation in eLearning

Classification and intent. Explanation pattern is based on Builder creational pattern [11]. Its intent is to help separate the construction of a complex object from its representation. Such a separation makes it possible to create different representations by the same construction process.

Motivation. Suppose an eLearning designer wants to develop an explanation generator that can generate explanations for different students. In general, current level of mastering the subject is different for different students at any given moment. That fact is reflected in the student model of each student. Novice students should get more general and easy explanations, while more complex and detailed explanations to more advanced students have to be provided [7]. The problem is that the number of possible explanations of the same topic or process is open-ended.

Using the Builder pattern provides a solution. The explanation generator in eLearning LMS could be designed with an *ExplanationBuilder*, an object that converts a specific knowledge level from the student model to an appropriate type of explanation, which is exposed in Figure 2. In this paper, *ExplanationBuilder* given in [7] is expanded and extended with helpful and misleading suggestions and hints, used for realization of *helpful* and *misleading* pedagogical agents.

The lecturer arranged and organized the appropriate explanations. Whenever the student requires an explanation, the explanation generator passes the request to the *ExplanationBuilder* object according to the student's knowledge level. Specialized explanation builders, like *EasyExplanationBuilder* or *Advanced-ExplanationBuilder*, are responsible for carrying out the request.

Structure. Figure 2 shows the general structure of the *Explain* pattern, based on Builder pattern. Unlike similar form, given in [7], *Explain* pattern is extended with helpful and misleading suggestions, hints, and clues.

Consequences. Using *Explanation* pattern lets designers vary a product's internal representation, e.g., the contents of the explanation. The pattern provides isolation of the code for representation from the code for construction. Construction of the product is a systematic process, and is under the director's control.

Known uses. Examples of using the *Explanation* pattern in Intelligent Tutoring Systems (ITS) design include different generators, such as explanation generator, exercise generator, and hint generator. In GET-BITS model [8], explanation generator is can construct explanations for a predefined set of users, which is configurable (e.g., beginners, midlevel, advanced, experts...) [9]. Hints for solving problems are generated in much the same way. In *Eon* tools, different contents are presented to the student during the teaching process depending on different Topic levels, which represent different aspects or uses for the topic (e.g., introduction, summary, teach, test, beginning, difficult,…) [10]. Extended *Explain* pattern is used in ***A**gent-based **E-l**earning **S**ystem* (***AE-lS***) [6].

Related patterns. *Builder* pattern is similar to the Abstract Factory pattern [11]. Explain pattern is based on *Expose the Process* [2].

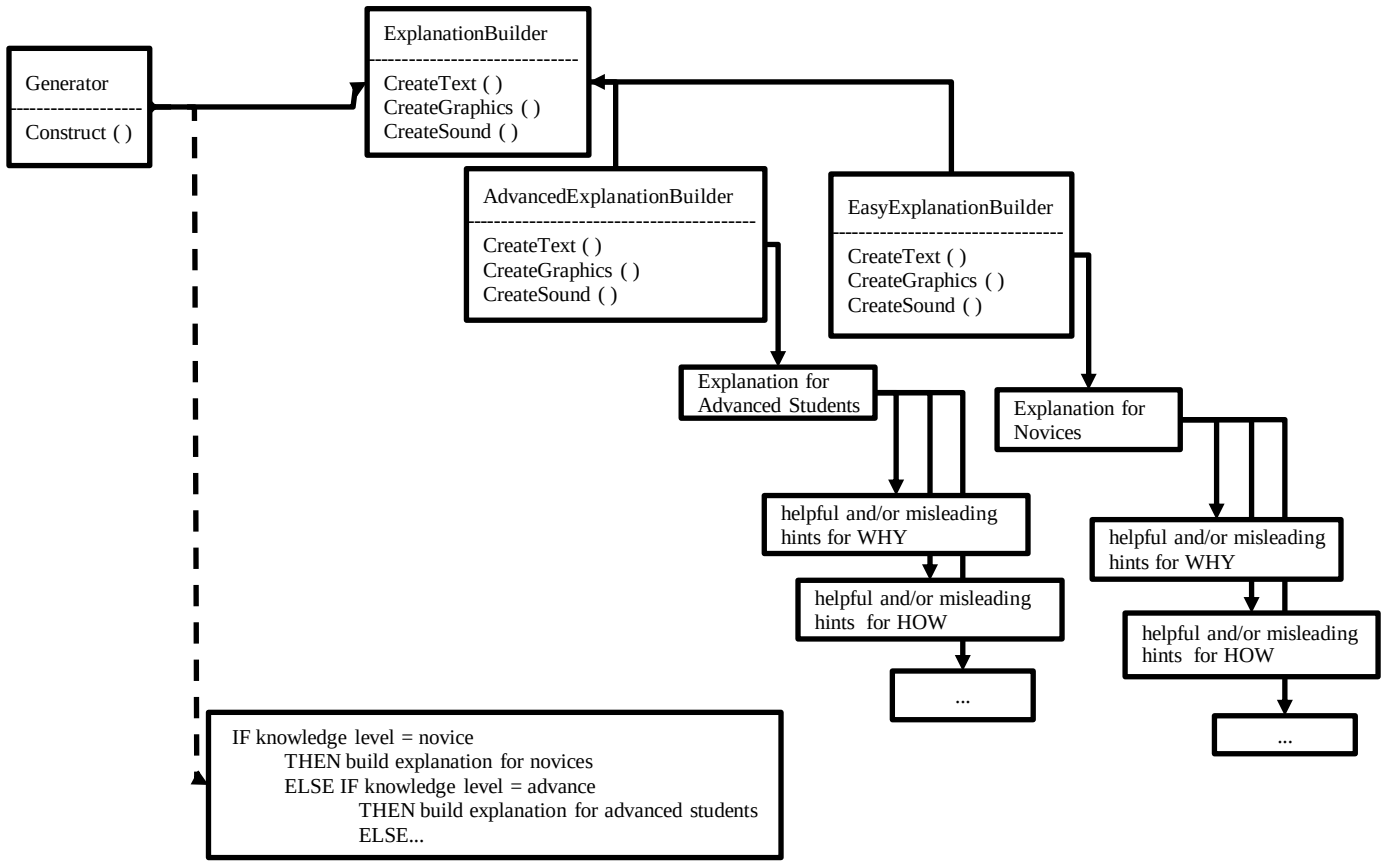


Fig. 2. Using the Builder pattern in designing explanation generator

C. Pedagogical Agents

The link between the student and the set of code completion tasks is provided in form of pedagogical agents. As noted earlier, two different types of pedagogical agents are used – one helpful, and one misleading in designing of ***A**gent-based **E-l**earning **S**ystem* (***AE-lS***).

As a crucial design step, both agents are hidden from the student behind the same interface, and take turns in interacting with the student at random time intervals. Therefore, the student is never sure with which agent he/she is interacting. The rationale behind this approach is straightforward: to motivate students not to trust the agent's hints blindly. Instead, they should critically analyze both the problem in question and the proposed hint, and, in the end, they them-selves decide on the proper solution.

In much of the scientific literature, as well as the actual software products, it is common to represent pedagogical agents as lifelike, animated characters. On the contrary, we feel that there is no real value in this approach. Primarily, many re-sources need to be put into designing and implementing a visually appealing character. However, although maybe "fun" to look at in the beginning, over the time the visual character and its built-in animations stand in the way of getting the job done. They distract the user/student from concentrating on the problem in question, and, in the extreme case, may negatively affect his/her willingness to use the system.

Pedagogical agents helpful and misleading are designed to increase student's productivity as primary goal. Consequently, no special attention for visual representations is considered. Purely, well-known characters from Office Assistant gallery, Clippy and Scribble, are used.

Both pedagogical agents are capable of adapting to each individual student. The agents track a set of information about the student, including his/her personal data, the ratio of correct and incorrect solutions to each code completion problem, and student's grade for each topic.

Based on the accumulated data, the agents can mediate if the student's success rate becomes unacceptable. For example, if the student gives to many wrong answers to questions regarding for loops, the pedagogical agent will recommend additional learning material or new examples, of course easier.

V. AE-LS EXAMPLE

Several important implementation requirements can be drawn from the functionality of ***AE-lS*** described earlier. For example, harvesting is a process that can and should be distributed and executed in parallel. Then, students should be able to interact with and use ***AE-lS*** through a web interface. Moreover, like all web-based systems, ***AE-lS*** should be resilient to hardware and software failures, malicious attacks, etc. Given these implementation requirements, and its popularity in developing software agents and multi-agent

systems, Java has been chosen as the implementation platform for ***AE-IS***.

A. Helpful and misleading hints

In order to provide the reader with a better insight into the evaluation of *AES*, some examples of the prepared code completion tasks are given next. The two given tasks are tailored to the topics on "*For Loops*" and "*Classes*" in Java, respectively. Helpful and misleading hints assigned to each task are also presented and discussed.

The task tailored to the topic on "*For Loops*" in Java requires the student to complete a program for calculating the first 10 members of the *Fibonacci* sequence. The skeleton program presented to students is shown in Figure 3 [6].

```
class Fib {
  public static void main(String[] args) {
    int[] f = new int[10];
    /* for loop goes here */
    print(f);
  }
}
```

Fig. 3. Code completion task related to for loops.

Based on this skeleton, a set of helpful and misleading hints for pedagogical agents have been prepared. The helpful agent uses the following set of hints:

H1. *for (int i = ?; i < 10; i++){}* "What should be the starting index? Remember that the first element of the *Fibonacci* sequence has the index 0, while the expression for calculating other elements is $f_i = f_{i-1} + f_{i-2}$"

H2. *for (int i = 0; i <= ?; i++){}* "What should be the ending index? Although you need 10 numbers, remember that the index of the first element is 0."

H3. *for (int i = 0; i < 10; ?){}* "Should you use *++i* or *i++* to modify the value of *i*? Remember that this modification is always executed at the end of the for loop"

The misleading pedagogical agent uses the following set of corresponding hints (Ivanovic et al., in press):

H4. *for (int i = ?; i < 10; i++){}* "What should be the starting index? Hint: the first element of the *Fibonacci* sequence is often denoted as f_0"

H5. *for (int i = 0; i <= ?; i++){}* "What should be the ending index? Hint: look at the initialization of the array *f* - how many elements does it have?"

H6. *for (int i = 0; i < 10; ?){}* "Should you use *++i* or *i++* to modify the value of *i*? Remember that the instruction *++i* first increases the value of *i*, and then uses the new value in an expression."

By suggesting that f_0 is the first element of the *Fibonacci* sequence in hint H4, the misleading agent tries to suggest the improper usage of 0 for the initial value of *i*. In the general expression $f_i = f_{i-1} + f_{i-2}$ this decision would cause the index to go out of the array bounds. Similarly, in hint H5, the agent suggests that the student should use 10 as the final value of *i* (note the expression *i<=?*), disregarding the fact that Java array indexes are 0-based. The final hint H6 is just trying to confuse the student (i.e. to check whether the topic "*For Loop*" mastered with comprehension or not), since obviously both *++i* and *i++* are correct.

The example given in Figure 3. is extended as following. Lecturer should pay special attention in assembling and incorporating the suitable and appropriate examples and tasks for learning and testing the student's knowledge. For example, instead to give the usual task for realizing the concept of array and the sum of some numbers (the use of topics "*For Loop*" and/or "*Recursion*" in problem solving), the following problem (task) is given to the students:

> ***"One mad scientist wants to make the chemical chain, made of plutonium and lead atoms. However, if two atoms of plutonium are side by side, the chain reaction and atomic explosion will be. How many of ways the safe chain could be constructed of the length N, if the mad scientist has N atoms of lead and N atoms of plutonium?",***

The goal of above task is to practice the recursive technique of programming and to compare their results with previously done. This problem is given instead the ordinary problem like:

> ***"Write Java method to realize the following mathematical function: $f_n = f_{n-1}+3$, $f_0=1$."***

The student's task is to write a method that calculates some function similar to the methods used in example for *Fibonacci* sequence. The helpful agent uses the following set of hints:

H7. Try to remember what we have done last two classes? Something about calculating "Fib… seq…" and "Rec… method."

H8. First, try to make model, i.e. appropriate series, of the sequence of the atoms.

H9. Use that initial value is 1. What is the next value? Find the connection between the first and the second value.

H10. Try $f_n = f_{n-1} + 3$, $f_0 = 1$

The misleading pedagogical agent uses the following set of corresponding hints:

H11. Try to remember what we have done last two classes? Something about calculating "*rectangle*…" and "*Rec*… method."

H12. It easy, you could try $f_i = f_{i-1} + f_{i-2}$. Yeah, that is model of the sequence of these atoms.

H13. Use that initial value is 0. What is the next value? Find the connection between the first and the second value.

H14. Get stuck? Try $f_n = f_{n-1} + 3$, $f_0 = 0$

VI. Pedagogical Patterns for Learning by Mistakes

Learning by mistakes is very fine teaching techniques or teaching method. In teaching Computer Science, Informatics, Information Technologies, and similar disciplines based on technique or technologies, and it is used very often. Joseph Bergin proposed couple of general Pedagogical Patterns, which are directly involved in learning by mistake method of learning, with special implications in usage of them in teaching Computer Science [12].

They are:

- ***Mistake*** - Students are asked to create an artifact such as a program or design that contains a specific error. Use of this pattern explicitly teaches students how to recognize and fix errors. We ask the student to explicitly make certain errors and then examine the consequences.
- ***Grade It Again Sam*** - To provide an environment in which students can safely make errors and learn from them, permit them to resubmit previous assignments for reassessment and an improved grade.

In addition, some other general Pedagogical pattern could be used to explore the method of learning by mistakes, with smaller modification [12]:

- ***Fixer Upper*** - the lecturer makes the errors and the students correct them.
- ***Test Tube*** – the lecturer ask for explorations. Here lecturer could ask for explorations of specific errors.

Couple of Composite Pedagogical Patterns could be used, like:

- ***Design-Do-Redo-Redo (DDRR)*** - pattern by Marcelo Jenkins [15], used in teaching Object-Oriented Programming (OOP) to senior students based on a multi-language approach. The idea is to teach OOP concepts such as encapsulation, abstraction, and polymorphism, independently of the OOP language used. To do that, a Design-Do-Redo-Redo (DDRR) pattern is used, in which students design an OOP solution to a programming assignment and then implement it in three different languages. They have to elaborate differences and possible errors.
- ***Design-Implement-Redesign-Re-implement (DIRR)*** – pattern by Steve Houk [16]. The pattern could be used to bridge the gap from an old paradigm to a new paradigm (from procedural to object-oriented), emphasizes common programmers mistakes when they tried to "compile" solutions form procedural point of view to object-oriented directly, for example.

In the next chapter, one new Pedagogical Patterns for using the learning by mistake method in teaching Computer Science will be presented.

VII. Pedagogical Pattern Wolf, Wolf, Mistake

Topic, which is taught, is divided into smaller pieces called subtopics or fragments. Fragments are introduced systematic using ***Spiral*** [12] or ***Semiotic Ladder*** [13] patterns. The goal of the topic is to show usage of these fragments in solving certain problems. After the whole material is presented, some examples of implementation these fragments (or the methods based on them) are shown to the students. They have active participation in constructing the solutions. At the end, an artifact such as a program, object and/or design, with a particular error has been realized. Lecturer knows that mistake is made, but say nothing about that. At the end of the class lecturer just says that all examples have to be tested and verified as homework assignment. Next time, lecturer asks students do they found something in their homework assignments. Lecturer is interested about their opinions on the correctness of the solution that he presented last time. Students should explain the nature and possible consequences of the error, if they were find the mistake at all. Lecturer just conducts the discussion. Using this form, students learn how to recognize specific errors of construction and design, as well as the importance of testing software.

In the rest of this Section, the definition of Pedagogical Pattern Wolf, Wolf, Mistake is presented.

Title: ***Wolf, Wolf, Mistake***

Problem/Issue: Novice students make mistakes in programming, design, and particularly in problem solving. Moreover, they are aware of that. Students "believe" that teacher is a person who always tells the truth, so they accept the facts and solutions without checking them. Moreover, the students take and accept some facts without checking the source of them, from Internet for example. Students often do not know how to interpret the error messages, or what to do to solve problems that are diagnosed. Debugging and Testing are an essential skill, whether done with a sophisticated debugger, or just by comparing actual outputs or results with expectations, as well as to have the whole picture of the problem and test properly the given solution from teacher.

Audience/Context: This is very applicable to the early stages of learning programming. Syntax and semantic errors are frequent and students need to become familiar with the messages produced by compilers and run-time systems. In addition, the students have to understand what these errors indicate about the program. More over this pattern is good in learning the students about importance of proper testing the solution in problem solving. The pattern could also be used in an analysis or design course in which certain specific, but common, errors could be made easily.

Forces: Students, make errors in problem solving, more than professionals and/or teachers. They are not prepared to see the whole picture, yet. Students do not accept easily the fact that testing the solution is very important.

Teachers usually help students to pass up possible errors in problem solving techniques, telling them about all cases that have to be considered, before the solution is constructed. Moreover, teachers know how to test the solution properly. Therefore, the students became passive, not active participant in learning process. They simply accept and memorize the solution, instead to construct it, in sense to create new knowledge of some topics.

Solution: Some carefully chosen example in problem solving technique is presented to the students. Teacher creates solution from the beginning (understanding of the problem) to the end (making the code). The given solution has certain (hidden) specific errors (usually a single error).

Teacher then asks students to carefully consider and explore given solution, to test it, and to find is it good or not.

When the students find the error, give them the chance to elaborate and discuss the cause and the consequences. Use ***Gold Star*** [12] for the reward.

If students do not find the error, tell them that the solution is not good in some cases. Give them extra time and/or some hints, trying to activate them. Repeat the process until the solution is found.

Discussion/Consequences / Implementation: Students become more familiar with testing the given solutions. They understand why the error occurs, and how to correct it. Discovering the error, students could learn to avoid making it. The goals are to teach students how to analyze the problem properly as well as importance of the testing.

Examples for the use of this pattern should be carefully prepared. Otherwise, if there are too many errors or mistakes are too obvious, contra-effect could be produced.

This pattern can be used in many situations. In design part of Software Engineering course, problem solving courses, Object-oriented courses, and the like, the pattern could be successful. Moreover, it can be used in introductory programming course.

Special Resources: The instructor simply needs knowledge of the problem he thought; therefore, he could hide the trap.

Related Patterns: ***Fixer Upper*** **[12]**, ***Test Tube*** **[12]** and ***Mistake*** **[12]**.

Example/Instances: This pattern could be used effectively in teaching some introductory CS course. If you wish to teach the students about importance of analyzing the boundary cases in program design, and why the testing software is not an easy job, you may use this pattern.

For example, the pattern was used in Basic of Computer Literacy course for non-professionals (like students with major in Geography) at the University of Novi Sad. Topic on data types and potential problems with them (such as division by zero for numbers, for example) was taught at the beginning of the course. After a while, branching and control structures were done, and their usage in solving some problems is presented. The students together with lecturer solve some problem using these branching and control structures. The lecturer conducted the output. Nevertheless, students, i.e. for the particular data entry the program could crush, do not see the "hidden" special case. They miss to observe the case, which leads in dividing by zero. This case lecturer "wisely" ignore in the analysis of the task. Next class, if the students still did not notice the mistake, and lecturer admitted her/his "sin", and explains the reason and consequences of mistake. Couple weeks later, students get the assignment very similar to previously, but in some other context. They all do the assignment without a single mistake.

In addition to those mentioned above, this pattern could be used effectively to teach students about pointers in languages like C or C++, by having them make all of the common pointer errors purposely. This particular use is somewhat dangerous on computers that have memory mapped I/O and unprotected operating systems. Both syntax and semantic errors can easily be explored using this pattern.

One exercise from an old book [14] was to write a program that produced every diagnostic mentioned in the manuals for a given (FORTRAN) compiler. This is, not surprisingly, very difficult to do. Impossible, for some compilers, as the documentation and the compiler are not parallel.

Contraindications: Do not use this pattern too often. You all know the fairy tale about a boy who cried wolf, wolf when there was none – everybody believed because he is a little boy, and they do not know to lie. He does it too many times, so when the wolf came, no one believed him. You could lose confidence and authority of experts in the eyes of students.

VIII. EXAMPLE OF PEDAGOGICAL PATTERN WOLF, WOLF, MISTAKE

"Our goal is to transform how children learn, what they learn, who they learn from." (Mitchel Resnick, A Media Lab for Kids: $27 Million from Isao Okawa Creates Center for Future Children at MIT, MIT News. November 18, 1998.)

Therefore, our starting points are:

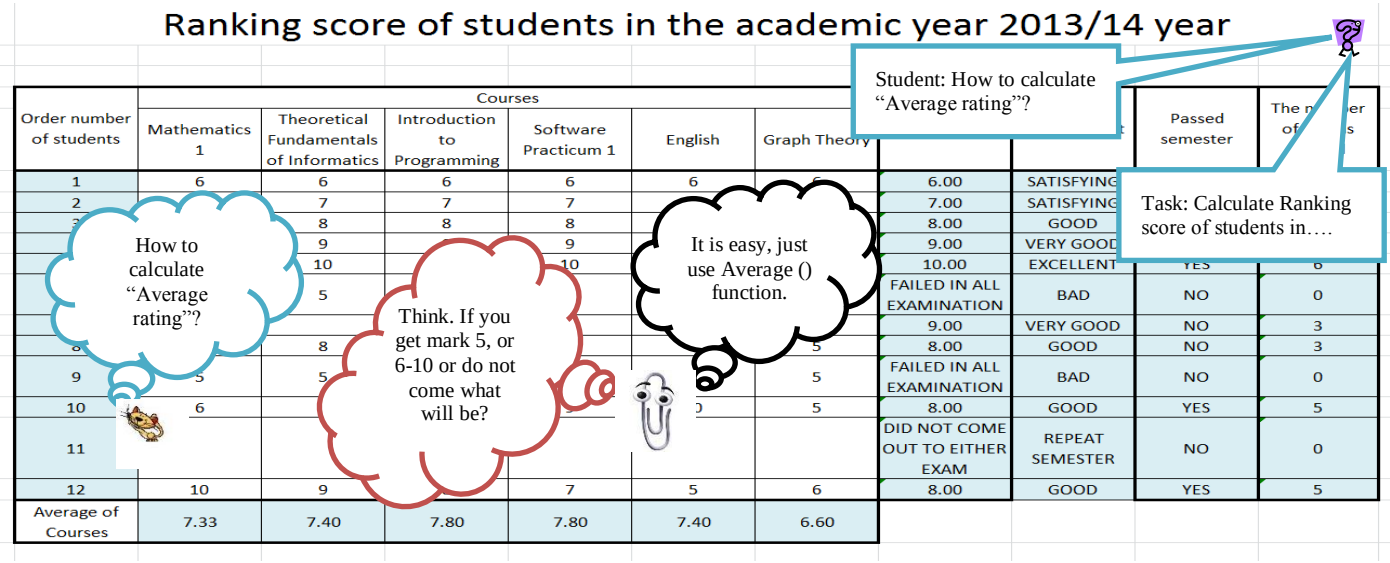
- We strongly believe that teaching is ART.
- Therefore, our first advice is to be a first-class artist on your stage (the classroom).
- It means, try to be different from others teachers in your environment, and engage your students to actively participate in lecture.
- Use a constructivist approach rather than objectivist in teaching.
- Use games and tools in teaching.

In addition, provide some home works for the students. For example, you finished classes about word processing in some course for computing literacy. After some time, give to the students your CV generating by Research Gate (for example), and ask them "How many times does my name appear in that document?"

Alternatively, novice students make mistakes in programming, design, and particularly in problem solving. Moreover, they are aware of that. Students "believe" that teacher is a person who always tells the truth, so they accept

the facts and solutions without checking them. Moreover, the students take and accept some facts without checking the source of them, from Internet for example. Students often do not know how to interpret the error messages, or what to do to solve problems that are diagnosed. Debugging and Testing are an essential skill, whether done with a sophisticated debugger, or just by comparing actual outputs or results with expectations, as well as to have the whole picture of the problem and test properly the given solution from teacher. For example, the pattern was used in Basic of Computer Literacy course for non-professionals (like students with major in Geography) at the University of Novi Sad. Topic on data types and potential problems with them (such as division by zero for numbers, for example) was taught at the beginning of the course. After a while, branching and control structures were done, and their usage in solving some problems is presented. The students together with lecturer solve some problem using these branching and control structures. The lecturer conducted the output. However, students, i.e. for the particular data entry the program could crush, do not see the "hidden" special case. They miss to observe the case that leads in dividing by zero. This case lecturer "wisely" ignore in the analysis of the task. Next class, students still did not notice the mistake, and lecturer admitted her/his "sin", and explains the reason and consequences of mistake "she/he made". The usage of pedagogical agents is provided, helpful and misleading. Therefore, the students could try to re-solve task (Figure 4.).

Couple weeks later, students get the assignment very similar to previously, but in some other context. They all do the assignment without a single mistake.



Order number of students	Courses									
	Mathematics 1	Theoretical Fundamentals of Informatics	Introduction to Programming	Software Practicum 1	English	Graph Theory			Passed semester	The number of
1	6	6	6	6	6	6	6.00	SATISFYING		
2		7	7	7			7.00	SATISFYING		
		8	8	8			8.00	GOOD		
		9		9			9.00	VERY GOOD		
		10		10			10.00	EXCELLENT	YES	6
		5					FAILED IN ALL EXAMINATION	BAD	NO	0
							9.00	VERY GOOD	NO	3
		8				5	8.00	GOOD	NO	3
9	5	5				5	FAILED IN ALL EXAMINATION	BAD	NO	0
10	6					5	8.00	GOOD	YES	5
11							DID NOT COME OUT TO EITHER EXAM	REPEAT SEMESTER	NO	0
12	10	9		7	5	6	8.00	GOOD	YES	5
Average of Courses	7.33	7.40	7.80	7.80	7.40	6.60				

Fig. 4. The usage of pedagogical agents is provided, helpful and misleading.

The lecturer have to provoke with the right questions/tasks, to determine how students have progressed and understand what you are teaches, as well to engage your students to participate in lecturing actively, with aim of pedagogical patterns and agents.

At the end, good lecturer do not forget to use good old methods in teaching, like use of some physical device, such as a toy, that has some of the characteristics of the concept being taught. For example, use the Frisbees™, to explain the concept of a parameter passing in function and the difference between value and reference parameters in function calls - elementary programming course.

IX. Conclusion and Further Work

The Pedagogical Pattern Wolf, Wolf, Mistake is described. The pattern could be systematized in category of General Pedagogical Patterns. The example of usage of the pattern is presented. In further work realization some more Pedagogical Patterns for learning by mistake method will be realized, like Blow-Up, Crash the System etc. Also the important part of teaching Computer Science (and other scientific fields), Explanation Pedagogical Patterns and will be investigated. In addition, the pedagogical approach "Gradual Improvement and Stepped Development with Fine-tuning" [17] based on pedagogical patterns will be further researched and developed, for teaching programming.

The agent technology has been recognized as a useful tool in a wide variety of domains. From the point of view of e-learning and tutoring systems, harvester and pedagogical agents are of the special interest. Some examples of using e-learning system named ***AE-IS*** that efficiently incorporate both harvester and pedagogical agents based on pedagogical patterns approach are given.

A more important functionality, however, is achieved by defining two new sub-types of pedagogical agents - helpful and misleading. As noted, the helpful pedagogical agent provides correct suggestions and hints for the problem in question. On the other hand, the misleading agent tries to steer the problem solving process in a wrong direction, by offering false suggestions and hints. The main motivation for this approach is to motivate students not to follow the agent's directions

blindly, but instead to analyze both the problem and the suggestions thoroughly, employ critical thinking, and, in the end, they themselves find the solution to the problem. According to our knowledge, none of the existing e-learning systems uses this kind of helpful and misleading pedagogical agents, in combination with pedagogical patterns.

The further work will be in two directions. First direction is definition of teaching agents. They help the teacher to build eLesson based on Constructivism. Constructivism offers a sharp contrast to teaching/learning [18]. First, the modern education is based on active student and student-centered teaching. Constructivism is a theory of learning based on the idea that knowledge is constructed by the knower based on mental activity. This approach is our contribution in replacing objectivistic learning theory at University of Novi Sad in teaching programming. In this view, students passively "absorb" programming elements, commands and structures presented by lecturer and documented in textbook, presentation, blackboard, etc. Teaching consists of transmitting sets of established facts, skills, and concepts to students. This is classical objectivistic approach in teaching.

The second direction is definition of "new pedagogy/didactics for teaching programming", called ePedagogy/eDidactics of programming, based on pedagogical patterns: ***Gradual Improvement***, ***Stepped Development*** and ***Fine-tuning*** [17], to promote Constructivism. If you examine the tables of contents of most eLearning systems, you find that the underlying educational philosophy is one of Objectivism. This theory holds that the student's mind is an empty slate that the lecturer/teacher/instructor fills up. The systems approach to this kind of eEducation has the creator of that system examine the subject to be taught, divide it up into small bits, sequence the bits in some logical order, and then put all students through the same process of learning the material in that order.

For example, eTextbooks (most of eLearning materials are some kind of electronic textbooks and called Tutorials) for learning elementary programming suggest that IF statements MUST come before LOOPING statements and so they contain chapters devoted to everything about selection, before anything is seen of repetition. These eLearning systems are reference works, not learning materials. The objectivist theory ignores the fact that such a methodology is deadly boring to most students. First, it forces them to "learn" things they already know. Second, it ignores any individual difference in learning style or preference.

Constructivist educational philosophy, on the other hand, views the student as knowledgeable and task driven. New things are learned by integrating them into what is already known and it is done primarily so that meaningful (to the person) tasks may be carried out.

At the end, the "future" of using computers in education is the last direction. Instructional computer programs (or the usage of computers in education) are being developed since the early '70s. Rapid development of Information Communication Technology, introduction of computers into schools, and daily use of computers by people of different vocation, education and age, has made education a very important field to researchers. Their main goals have been to develop programs that can teach humans and to achieve individualization of the educational process. The methods and techniques of Artificial Intelligence have been successfully used in these systems, since the end of last century. Hierarchical modeling, interoperable and reusable software components, and ontology are modeling techniques that have only recently penetrated into the eLearning. In addition, these methods are used in new "field" called "eEducation", a new approach to education with the help of Information and Communication Technologies and Computer Science. The following questions have to be answered:

- Could we described "eEducation" = "eLearning" + "eTeaching", by this "simple" equation? Alternatively, do we need more "+"?
- Are we all (researchers, teachers and students) have succeeded in eEducation (eLearning) so far? Do "users" of eEducation (eLearning) systems are "better" than traditional students are, in a since of learning gain?
- Do we have right pedagogy (teaching methods/strategies) for eEducation (eLearning)?
- Do we have right learning strategies (models/theories) for eEducation (eLearning)?
- At the end, what is the future of eEducation (eLearning)?

REFERENCES

[1] Richland, L. E., Kornell, N. and Kao, L. S., "The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning?" Journal of Experimental Psychology: Applied, 2009, Vol. 15, No. 3, 243-257.

[2] Bergin, J., Eckstein, J., Manns, M. L., Sharp, H., Maraquardt, K., Chandler, J., Sipos, M., Völter, M. and Willingford E., "Pedagogical Patterns - The Pedagogical Patterns Project", in Bergin J. (Ed.) *Pedagogical Patterns: Advice fir Educators*, Joseph Bergin Software Tools, 2012. (ISBN: 978-1-4791718-2-8)

[3] De la Prieta, F. and Gil A. B., "A Multi-agent System that Searches for Learning Objects in Heterogeneous Repositories," in *Proc. PAAMS Special sessions and workshops: Trends in Practical Applications of Agents and Multiagent Systems*, 8th International Conference on Practical Applications of Agents and Multiagent Systems, Salamanca, Spain, 2010, pp. 355-362.

[4] Haake, M. and Gulz, A., "Visual Stereotypes and Virtual Pedagogical Agents," *Educational Technology & Society*, vol. 11 no. 4, pp. 1-15, Oct. 2008.

[5] Heller, B. and Procter M., "Animated Pedagogical Agents and Immersive Worlds: Two Worlds Colliding," in *Emerging Technologies in Distance Education*, G. Veletsianos (Ed.), Athabasca, Canada: AU Press, 2010, ch. 16, pp. 301-316.

[6] Ivanovic, M., Mitrovic, D., Budimac, Z., Vesin, B. and Jerinic, Lj., "Different Roles of Agents in Personalized Learning Environments," In *Proc. of the 10th International Conference on Web-Based Learning - ICWL 2011*, Hong Kong, Dec. 8-10, 2011.

[7] Jerinic, Lj. and Devedzic, V., OBOA Model of Explanation Module in Intelligent Tutoring Shell. SIGCSE Bulletin, Vol. 29, Number 3, September, ACM PRESS, 133-135.

[8] Devedzic, V. and Jerinic, Lj., "Knowledge Representation for Intelligent Tutoring Systems: The GET-BITS Model", In: du Boulay, B., Mizoguchi, R. (Eds.) Artificial Intelligence in Education, IOS Press, Amsterdam / OHM Ohmsha, Tokyo, 1997, 63-70.

[9] Devedzic, V. and Jerinic, Lj., "Explanation in Intelligent Tutoring Systems", *Bulletins for Applied Mathematics*, 1196/96, 1996, 183-192.

[10] Jerinic, Lj. and Devedzic, V., An object oriented shell for intelligent tutoring lessons. Lecture Notes in Computer Science Vol. 1108, 1996, 69-77.

[11] Gamma, E., Helm, R., Johnson, R. and Vlissides, J., "Design Patterns: Elements of Reusable Object-Oriented Software", Addison-Wesley, Reading, MA, 1994.

[12] Bergin, J.: Fourteen Pedagogical Patterns. In M. Devos and A. Rüping (Eds.): Proceedings of the 5th European Conference on Pattern Languages of Programms (EuroPLoP '2000), Irsee, Germany, July 5-9, 2000. UVK - Universitaetsverlag Konstanz 2001 ISBN 978-3-87940-775-0, pp. 1-49, 2000.

[13] Kaasbøll, J. J., "Exploring didactic models for programming", Tapir, pp. 195-203, 1998.

[14] Teague, R., "Computing Problems for FORTRAN Solution", Canfield Press, 1972.

[15] Jenkins, M., "Pedagogical Pattern #13: Design-Do-Redo-Redo (DDRR) Pattern", in Bergin J. (Ed.) *Pedagogical Patterns: Advice fir Educators*, Joseph Bergin Software Tools, 2012. (ISBN: 978-1-4791718-2-8)

[16] Houk, S., "Design-Implement-Redesign-Re-implement (DIRR) – Pattern", in Bergin J. (Ed.) *Pedagogical Patterns: Advice fir Educators*, Joseph Bergin Software Tools, 2012. (ISBN: 978-1-4791718-2-8)

[17] Jerinic, Lj. "Pedagogical Approach 'Gradual Improvement and Stepped Development with Fine-tuning' in Teaching Programming", in print.

[18] Jonassen, D., "Objectivism vs. Constructivism", Educational Technology Research and Development, 39(3), pp. 5-14, 1991.

Development Process Patterns for Distributed Onshore/Offshore Software Projects

Ravinder Singh
AVP, JP Morgan Chase & Co
Research Scholar,
Department of Informatics,
King's College, London, UK

Dr. Kevin Lano
Reader
Department of Informatics,
King's College, London, UK

***Abstract*—the globalisation of the commercial world, and the use of distributed working practices (Offshore/ onshore/ near-shore) has increased dramatically with the improvement of information and communication technologies. Many organisations, especially those that operate within knowledge intensive industries, have turned to distributed work arrangements to facilitate information exchange and provide competitive advantage in terms of cost and quicker delivery of the solutions. The information and communication technologies (ICT) must be able to provide services similar to face-to-face conditions. Additional organisations functions must be enhanced to overcome the shortcomings of ICT and also to compensate for time gaps, cultural differences, and distributed team work. Our proposed model identifies four key work models or patterns that affect the operation of distributed work arrangements, and we also propose guidelines for managing distributed work efficiently and effectively**.

Keywords*—Distributed work; Onshore; Offshore; Software; IT projects; Programme and project Management

I. INTRODUCTION

People and organisations have been communicating and managing work over long-distances and multiple countries since ancient times also. Earlier, such distributed work and exchange of information was achieved by the physical travel of people, which made the flow of information slow and coordinating the work tedious and also costly.

Distributed environment of projects in the present multinational organisations gives rise to more complexities in all areas of project management. Therefore standard project management methodologies have to be enhanced to meet diverse requirements from various stakeholders. The studies showed that distributed work environment has its own challenges and advantages. The challenges could be such as managing different time zones, cultural differences, virtual communication environments and costs associated with them, and many more. The advantages could be in terms of providing good quality projects at lower cost. This requires proper documentations, setup the correct expectations, managing various stakeholders and also managing the cross cultural issues effectively and efficiently. The conflict resolution criterion and transparent communication is the key to success in global scenarios and managing successful projects.

Previous research had been focusing on different aspects of the program and project management such as study of models and framework, empirical, and statistical studies. The studies had been conducted in different industry sectors but most of the research has been in the software and IT industry as given in the following table.

Varied results from the work put organisations in difficult situations for the standardisations of processes to implement distributed work environments. Previous researchers have implied that this may be due to the lack of well-established framework for distributed work environments. One of the solutions could be to use the standard organisational theories to overcome the problems of distributed work environments. Even these theories are not sufficient to address the issues of the distributed work environment.

This paper proposes a new set of frameworks and identifies five models for using the distributed work more efficiently and effectively. The work highlights the use of various models and the conditions for its use. This work also put forwards different guidelines for helping to complete the distributed work in a more organised manner. These models are then applied to two organisations to see their impact on the overall performance of teams.

This paper introduces the various models available for distributing work between a customer site and the delivery/development centre (DC) network. These models are applicable for moving work to onshore/ offshore/ near-shore DCs. However, moving work offshore introduces additional risks that are explained in more detail in Risk Management Guidelines for Distributed Work.

II. DISTRIBUTED WORK APPROACHES

The details about the four model (customer-centric, DC-centric, multi-centre and tailored) is explained below along with a brief overview and the main characteristics of each. Various work models, when to use a particular model and application of each model is also discussed.

Ref. No.	Category/ Topic	Study Description/ Method/ Argument/ Theoretical Approach	Results, gaps, and Conclusions
	Project Management in Global Distributed Environment	Managing projects in global distributed has its own challenges [1-24]. Researchers had explored use of different methodologies, techniques, tools for managing distributed projects from standard processes to incentive based approaches.	
1.		With the exponential growth of communication technologies and information systems, the globalisation of the commercial world has also increased significantly.	This research paper highlighted that in order to increase efficiency, productivity, quality and cost effectiveness, organisations are going for outsourcing and distributing their wok globally.
2.		This research study described the importance of software requirement specification (SRS) document to the success of global software projects. The authors discussed various difficulties in creating a standard SRS as companies have their own methods of creating such documents.	The authors studied how Capgemini overcame the issue of creating standard SRS by using specification patterns so as to create synergy among the global teams.
3.		The significance of knowledge sharing among global teams and stakeholders and how it can be addressed by mature processes and tools is highlighted in this study. There will be lesser readjustment required if the processes, methods and tools are used enterprise wide.	The authors proposed that enterprise wide software should be used for project assurance, quality and knowledge sharing. The software would help provide timely information, data and visibility for the preventive and corrective actions to be taken for better execution of the project.
4.		This study described the team structure for successful completion of offshore projects. The authors studied two types of structures for offshore teams and highlighted the problems faced by managers for changing the team structure and organisation model.	The paper proposed that changes have to be done to the existing structure for successful global operations. The team structures for managing offshore teams for various phases of the project and the reporting structure has to be managed keeping tin to account various time zone issues, cultural issues and skills availability.
5.		A framework for managing risks in global software projects is proposed in this research paper. The integrated framework had been created for distributed projects based on various parameters and requirements of global environment.	The framework proposed the use of various communication channels, different set of development environment for different needs/ requirements of the stakeholders and projects. The flow chart could also help to provide better information across the organisation.
6.		This research studied the impact of communication media like email, messaging, phone etc. on the conflict resolution in global teams. The authors tried to evaluate which could be the best sequence or combination of media tools for communication for resolving the conflicts.	The study showed how the cross cultural issues, different communication channels, time zone management had to be taken into account for managing global teams/ people effectively. The process for conflict management has to be robust and transparent so that the conflicts can be controlled/ resolved in an efficient manner.
7..		In this study, authors tried to analyse the global development projects using framework so as to overcome various issues in the distributed projects. The authors tried to study the processes used by various organisations to manage the distributed projects efficiently and effectively, and maximise the benefits of onshore-offshore delivery.	The paper showed different models and frameworks used by global organisations to manage the distributed projects successfully. Various activities can be distributed offshore/ near-shore or onshore and also the life cycle divided among them for maximising the benefits.
8..		This research studied different communicating media and its application the global agile software development projects.	The authors found that instant messaging is a good substitute tool for face to face communication and email is good tool for wider and enterprise wide information sharing.
9.		This research paper proposed predicting the outcome of global software development projects with the application of analytical modelling. The analytical models are parameterised to accommodate the single-site or multi-sites, team sizes, skills levels, expertise, availability, and support level etc.	The paper suggested various types of models for distributing various phases/ stages between offshore and onshore sites.
10.		This research study described the processes for managing a multi-site software development project is complex and requires a very good collaboration among teams.	The study suggested management of multi-site projects can be improved using networked virtual environment which allows for better communication, familiarity, sharing, mentoring, faith and faster resolution of conflicts.
11.		This research studied the growth of teams in distributed software development projects. The authors had tried to study the growth of teams in terms of expertise, communication skills, economic impact and working conditions.	The study described the communication channels, skills and the impact of virtual communication techniques for successful management of teams and projects in global environment. The better the economic and working condition, the better would be the team morale and more successful project management.
12.		This research paper explained that the “Distributed Work” is basically a	The research highlighted the importance of

		number of different work provisions. Since the teams are distributed globally, and are separated by time zones, the managers have to rely heavily on the availability and efficiency of communications tools and information systems.	communication tools and information systems for successful management of global teams and projects.
13.		Use of incentive based theories to the distributed work environments is described in this research paper. The paper endeavours to address two subjects; firstly, to understand the effect of incentives on the worker's choice for using distributed work environment, and secondly the collaboration of multiple incentives or disincentives across organisation, groups or individuals. This paper also looks into motives as to why people always prefer to take up distributed work environment. The theory of incentive is applied to two organisations to understand the behaviour and pattern.	The research suggested that people prefer distributed work environment because of flexibility, incentives, and availability. The disincentives are managing different time zones and culture. The study showed that incentives highly influence the working of people and opting for distributed working. It also highlighted that work life balance is one of the main criterion for people for remote/ home working.
14.		This research paper studied as why organisations choose for distributed work environment. The research was conducted to understand the use of distributed work environments in terms of costs, efficiency and productivity, motivation of employees, and impact on the group's outputs.	The research suggested that the use of distributed work environments is to mainly reduce the costs, improve efficiency and productivity, motivate employees, and impacting the group outputs positively.
15-18		Even though there is clear impact on the employees for the work-life balance, more flexibility but there are conflicting observations made which are owing to more distractions at home which results in increased stress.	These papers showed that remote working, home working or flexible working is able to provide better work life balance but at the same time needs more planning as it could also lead to more distractions at home and less work. The employees have to manage themselves more efficiently to be more productive. Organisations provide hot-desk facilities to save on cost of space and also improve its travel carbon footprint.
19.		This research paper defined knowledge intensive firms as those that "offer to the market the use of fairly sophisticated knowledge or knowledge-based products". Knowledge intensive firms can be divided into professional service, and research & development firms such as engineering and law firms or pharmaceutical companies. Knowledge intensive firms differ from other types of organisations through the organisation's massive reliance on the intellectual skills of its employees to carry out its core functions.	Although many of the problems and barriers to distributed work are not unique to knowledge intensive firms, the sophisticated nature of the knowledge these firms typically deal in has the potential to magnify these problems. This report focuses on the interaction of individuals and teams within knowledge intensive firms and the ways that they interact and perform under distributed work arrangements.
20.		This research defined a virtual team as "groups of people employed in a shared task while geographically separated and reliant on electronic forms of communication".	The research paper compared various factors such as telephonic conferences, video conferences, e-mails, time zones, and for managing virtual teams. The virtual communication tools are important and also people should be sensitive to the cultural communication styles and language used in communication to overcome misunderstandings and reduce communication gap.
21.		The paper defines the term remote resourcing as "carrying out work in an office remote from the point where a project is principally delivered". The report defines remote resourcing when virtual communication tools are used and teams are distributed at one or more sites in different geographical locations.	These terms essentially describe interactions between people separated by physical distance who perform most of their work through communication technology. Within the body of this report the term distributed work is used to represent this concept. The dynamic changes to the project are handled more effectively when the team is at one place and long-term projects can get greater benefits from remote teams or by distributed working.
22.		The research paper discusses that distributed work covers many alternative methods of work which include satellite offices, flexible work arrangements, telecommuting and global collaborative teams.	The paper describes that distributed work could be defined in many different ways. The distributed teams could use different ways of working from flexible home working to offshore, onshore or near-shore arrangements. The paper highlighted that distributed teams and working are often used to reduce overall cost and improve services.
23.		This paper describes various issues and problems faced by distributed work faces which are similar to all the issues and problems that normal collocated group's face, with the added complexity of workers being based at locations remote from each other, be it in the next room or in another country The inclusion of IT as a required element of many definitions reflects the importance of ICT as a replacement media to mimic the communicative and collaborative qualities inherent in collocated work groups.	This paper highlighted that distributed work faces many more problems in addition to the normal projects at one site. The projects and teams distributed in different locations brings in the importance of good communication media and skills, cross cultural issues and management, time zone management, and clear understanding of the stakeholders' expectations. The project documentation has to be detailed and

			shared with all teams highlighting various milestones and deliverables and also giving details of communication requirements.
24.		This research explained that small and medium enterprises (SMEs) are also facing huge competition due to globalisation of economies and easier availability of cheaper and good quality products, services across the world.	This paper highlighted that in order to stay ahead of the competition and technology SMEs should focus on to e-collaborations through project management approach. This will ensure them structured processes, better visibility for managing the full life cycle of the project and giving them better monitoring and control of project execution.

III. Customer-Centric Model

With the use of this model the majority of the work is completed at the customer site, and the detailed design, build, and component tests are done at the delivery/development centre.

The customer team transfers the well codified tasks to the delivery centre to be executed with the most discipline and rigor. This distribution model can be used for both onshore and offshore delivery centres and may have to be adapted to suit specific constraints of the project and stakeholders.

Main characteristics of the Customer-centric model are as follows:

- The most basic model, suitable for first time users
- Moderate cost savings
- Moderate risk
- Suitable for all project sizes
- Limits cost savings because only a small portion of the life cycle is completed at the delivery centre
- May not be suitable for development of components that involve a high degree of communication with the customer (e.g., UI, data manipulations, etc.)
- May not be suitable for development of new/complex applications

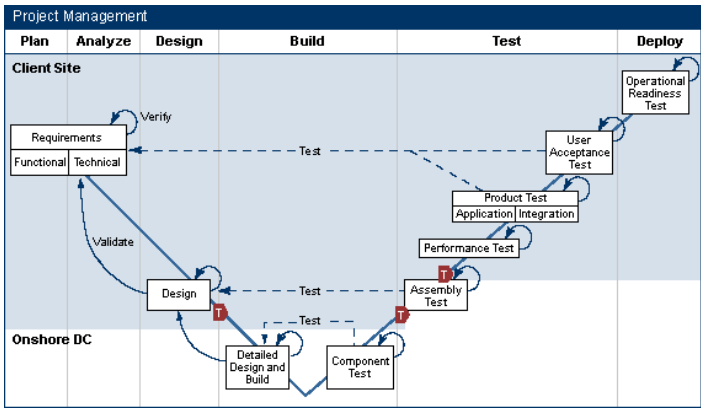


Fig. 1. Customer-centric Distribution Model

This model represents a minimum amount of risk from the long-distance cross-site communication. The physical distance has less of an effect on communication than significant time zone differences, but Time zone differences can prevent project team members from communicating with each other in real time. Even though this is the most basic distribution work environment, it may still be a perfect model to execute "forever" depending on the stakeholder expectations.

Benefits

- Simple, stable, and repetitive processes. Only a small portion of the development life cycle is executed at the delivery centre. Transition points control the interaction between customer and the delivery centre sites. Also, formal and informal communication ensures a proper flow of information.
- Robust and scalable. The process' simple design gives the delivery centre site these characteristics. This will achieve cost-savings.
- Minimal communication. The delivery centre site's communication is between the design and build teams and rarely involves the customer. The low amount of communication is because of the formal and specific design deliverables that are less open to interpretation than requirements.
- Works well with offshore centres. Due to all previously listed characteristics (repeatability, scalability, and robustness); this distribution model works well with offshore centres.

Cross-site liaisons ensure a smooth issue resolution process.

Drawbacks

- Since this model limits the types of tasks which can be done at delivery/ development centres, therefore cost-savings which can be realised are also limited.
- Assembly tests may not be fully conducted at delivery/development centre when an application comprises cross-platform assemblies of components and these components are developed using separate toolsets.

Applications

- This model is particularly desirable for custom-based or packaged solutions that require a pool of skilled programmers producing large-scale applications.
- When planning to work with an offshore centre, use a nearby onshore centre as an intermediary as this will save time and effort during project planning and the project execution phases. Onshore centres should have

more application analysis and business skills than their offshore counterparts.

- Impact of required levels of communication.

- Since this model uses minimum communication with the customer, therefore it may not be suitable to developing the application components that require a high degree of communication with the customer (e.g., UI components, reports, integration etc.).
- This model may not be suitable for developing application components that fit into a new application architecture, as it may require a high-level of communication with the design team. This can be mitigated by having the technical architecture development team at the delivery centre. For developing a new architecture, completing it as "Release 0" at the customer site will reduce the risk.

- Transition of the application back to the onshore team and whether this occurs before or after assembly test (indicated by the red transition points) needs to be carefully considered. Transition prior to assembly test means a change in team and ownership, but may be required due to technical testing constraints (e.g., cross-platform environments) or contractual obligations (e.g., only delivering one part of the application). However, where possible, execution of assembly test is more effectively performed by the development team prior to any significant handover or transition to another organization (e.g., the formal onshore test team).

There are circumstances where even the most basic distribution models cannot be executed and require all tasks to execute at the customer sites. For instance, if the customer is uncomfortable or unwilling to see part of the effort executed at a delivery/development centre or has a particular environment, the delivery centre personnel can work at the customer site.

IV. DC-Centric Model

In this model, most of the work is done at the delivery/development centre. The customer site completes only requirements gathering/analysis and user acceptance testing. DC-centric model characteristics include the following:

- Moderate cost savings when applied with an onshore centre
- Significant cost savings when applied with an offshore centre
- Low risks when applied with a onshore centre
- Increased risk with the distance and time zone differences between the customer and delivery centre sites
- Suitable for a wide variety of applications
- Suitable for use with all project sizes
- Requires a higher maturity DC and team experienced with multi-site projects to execute

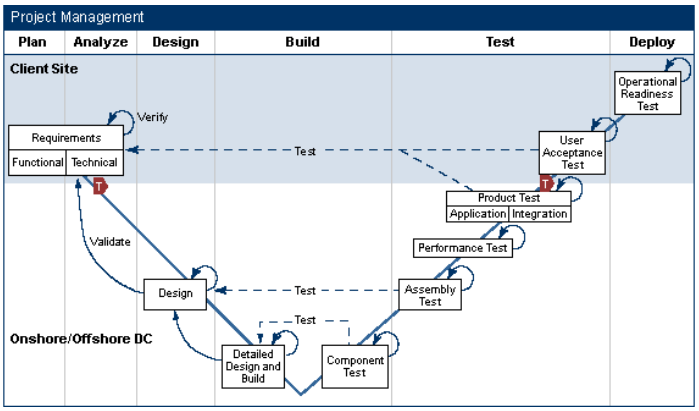


Fig. 2. DC-centric Distribution Model

DC-centric distribution model enables significant cost-savings at low-to-moderate risk levels because of the task distribution. This is the predominant distribution model used for custom development by the onshore/offshore/near-shore centres today may have to be customised to the suit specific constraints of the project and stakeholders.

This model can be used only when the customer team is experienced in delivering projects with offshore centres, and the offshore centre is relatively mature (e.g. CMMI Level-2 or higher, Six sigma, etc.) and has demonstrated expertise in the project management, technologies and applications. In order to reduce risk, start the project with a more basic approach, i.e., the Customer-centric Distribution model, and then progressively migrate additional activities offshore. The desired end-state is best achievable over a period of time.

Characteristics

- This model requires higher levels of communication between the sites than the Customer-centric Distribution model. The key transition point between the sites in this distribution model lies between analysis and design, while in the Customer-centric model focuses between design and detailed design:

- In a typical situation, Transition Point Overview results in a higher level of communication between the sites because it may involve communication and resolution of issues with the customer and the set-up is less tolerant to delays caused by distance and time zone separation.
- Application design deliverables are easier to specify than application requirements Transition Point Overview to a sufficient level of detail and without (or with less) ambiguity. This makes the application design deliverables less prone to misinterpretation. Detailed standards exist for specifying the design, while requirements are typically defined more generally and are open to broad interpretation.
- Since this model requires higher levels of communication, it will work well with delivery centres in close time zone proximity to the customer sites. Significant time zone differences will make it difficult for team members to communicate synchronously.

- Engagements based on established offerings and/or assets are particularly well suited to this model, since there are fewer risks related to miscommunication when using stable technologies, environments, and processes.
- This may be the predominant model for working with onshore centres.

Benefits

- This model will enable the realization of maximum cost-savings, as most development tasks are completed at the delivery/ development centre with a more cost-efficient workforce, standard repeatable processes, application-specific methodologies and job aids, reusable assets, etc.
- Since this model was previously used at onshore centres, significant processes, experts, and procedures can be used for the effective management.

Drawbacks

- The distance and time zone differences between the customer and delivery/development centres increases risk.
- This model requires mature (e.g. CMMI, Six Sigma, ISO etc.) and experienced offshore centres to work successfully.

Applications

- Address the risks through various risk mitigation strategies when applying this distribution model with offshore centres:
  - To reduce the communication gap and reduce the rework activities, investment is required communication infrastructure (e.g., internet connectivity, configuration management tools, video conferencing, etc.).
  - Customer can build and invest in the communication technologies at site only if the project is long-term to recover the cost. Otherwise customer can use third party service providers to meet short term goals.
  - An onshore or near-shore centre as an intermediary may be used when using an offshore centre to reduce start-up costs and to reduce the issues related to offshore development.

V. TAILORED MODEL

With the maturity of customer team and offshore centre, the location of the individual tasks is determined by the cost/benefits/risk analysis. This distribution of tasks at individual levels poses more complexity but it provides optimisation of cost/benefit/risk. The tailored model characteristics are as follows:

- Optimal and balanced in terms of costs, benefits, and risks
- Suitable for all project sizes
- Requires experts and maturity of processes to plan and execute

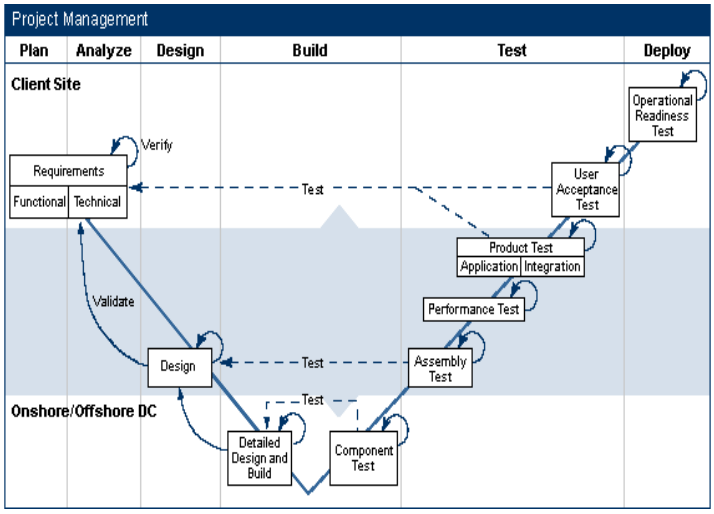


Fig. 3. Tailored Distribution Model

Creating a Tailored Distribution Model

This method requires experience and help from a delivery centre expert who is familiar with cost-risk-benefits analysis of multi-site development in offshore centres.

The method works with a two-dimensional matrix where functional areas are derived from the application requirements. The horizontal dimension corresponds to the major phases of work, such as analysis, design, component test, etc. The vertical dimension corresponds to the functional areas within the application, such as IPO, Billing, and Account Management.

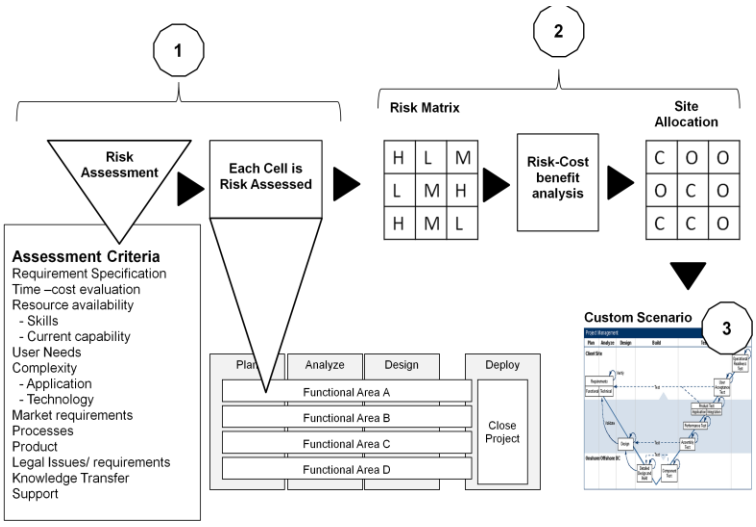


Fig. 4. Creating a Tailored Distribution Model

There are three major steps in the Tailored model:

- Assess individual criteria for each matrix, and determine the aggregate risk for a given stage/phase of work for a given functional area. The resulting

aggregate value of High, Medium, or Low indicates the risk for executing a given stage of work for a given functional area offshore.

- Apply cost-risk-benefit analysis to each matrix cell to determine whether to execute a given stage at the onshore or the offshore centre. Consider factors such as skill availability, cost, potential knowledge transfer, etc. The result of this step is a site assignment matrix, with each cell containing a designation "C" Customer site, "N" Near-shore/onshore centre, or "O" offshore centre.
- The resulting matrix can be used to plan the work/ tasks.

Although the process seems simple and straightforward, it will require experts and maturity to conduct cost-risk-benefits analysis.

VI. Multi-Centre Model

In this model, the work is distributed across at least three different sites: the customer site, the onshore/near-shore centre site, and the offshore site. The requirements gathering and analysis and the user acceptance testing are completed at the customer site. The rest of the work is shared between the onshore/near-shore and the offshore centres.

This model is able to provide the benefits of the both the DC-centric and customer-centric models. Greater cost-savings are achieved by using the offshore centre and the risk is reduced because the customer team works closely with onshore/near-shore centre.

The use of this model is on the rise, and it will be a predominant approach in the future, particularly for packaged-based development. Multi-centre model characteristics include the following:

- Combines benefits of the other two models
- Model of choice for packaged-based development

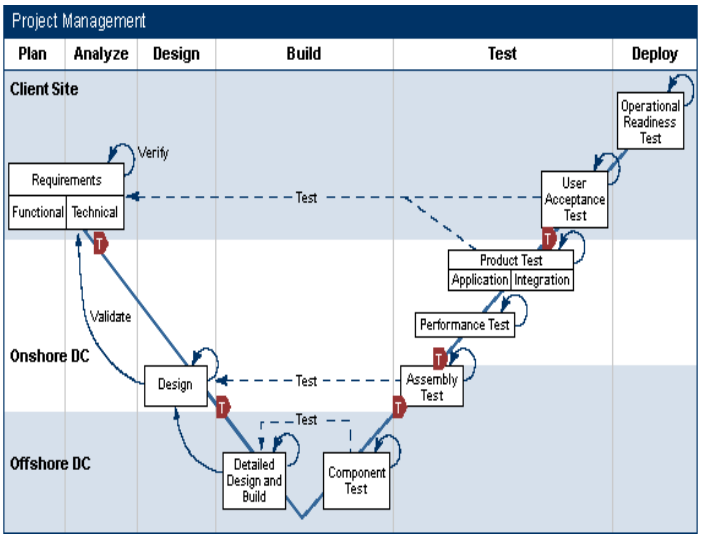


Fig. 5. Multi-centre Distribution Model

- This model is most suitable for medium and large-scale projects, as it typically involves significant start-up costs (training, infrastructure, and knowledge transfer). However, this model can work for smaller projects if they can use an existing facility or run from the same delivery/development centre.
- This model requires experts and maturity of team. Typically, the customer site team drives the business requirements while the onshore/near-shore centre drives the technical delivery work. The onshore centre also acts as a hub between the customer and the offshore sites.
- This model may not be fully suitable for projects with well established application and technology architectures because they do not require the onshore/ near-shore centre to act as a liaison between the customer site and the offshore centre. In such a case, DC-Centric model may be more suitable.
- If projects are based on a new architecture, consider a different distribution model. The complexities of dealing with three different sites are magnified by the complexities associated with managing the new architecture development.
- This model is particularly well suited for packaged software delivery:
  - The model will work well with moving defined work offshore and keeping more difficult and less defined work onshore/ near-shore.
  - Working through an onshore/ near-shore centre also helps overcome language barriers, time zone differences, etc.,

Benefits

- Cost savings. This model's scalability can help achieve greater savings for projects with a large build component, while shielding the customer site team from the exposure to the complexities of dealing with remote delivery centres. By using the delivery/development centre to complete more tasks, additional savings can be achieved.
- Lower risk. The risk is lowered as the near-shore/ onshore centre manages the tasks that require higher levels of communication with the customer site team (e.g., UI design, functional design etc.). This mitigates the risk of communication gap and delays.
- Higher quality. The near-shore/ onshore centre ensures the errors are corrected before the customer site receives the build components. The centre does not necessarily inspect the quality, but it will be able to facilitate the transition smoothly.

Drawbacks

- Since it involves significant start-up costs. It also requires an experienced team for execution, this model is suitable only for medium and large projects.

VII. Distribution Approach Versus Maturity/ Experience

The diagram below depicts the relationship between the ability to execute higher complexity distribution approaches and the organizational maturity. The organizational maturity combines two notions:

- Delivery centre maturity. This may be referred as CMMI level rating, Six Sigma, ISO certification etc. attained by the delivery/development centre.
- Customer site team maturity. This is the customer site team's experience with multi-site project execution. This is often reflected in the number (percentage) of the management and development people who previously worked on a multi-site project, involving a delivery/development centre.

The graph of distribution approach vs. experience/maturity follows the S-curve, with use shifting from customer-centric scenario to DC-centric scenario.

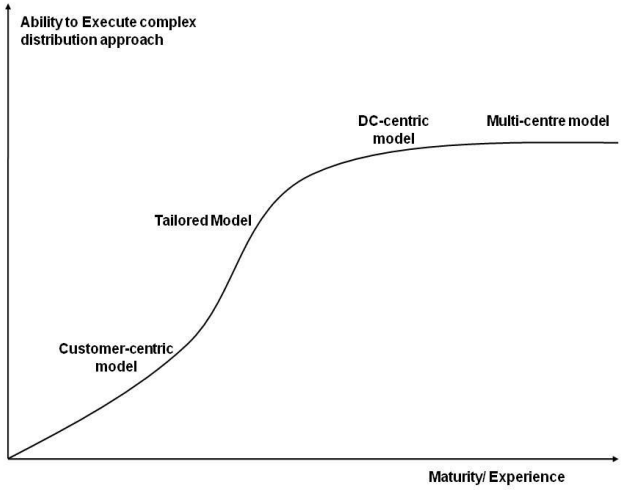


Fig. 6. Distribution Approach vs. Maturity/ Experience

The implications and considerations for the amount of experience working with a distribution approach include the following:

- The customer teams new to the multi-site may want to start with the customer-centric model.
- The customer team's maturity can be influenced by acquiring people experienced with multi-site development from a delivery centre for the team. Start with the DC-centric model if you have the right people.
- The primary considerations for selecting a specific model are listed in the table. There are situations to execute the customer-centric model long-term, regardless of experience or delivery centre maturity.
- Longer-term engagements must consider the potential for starting with a more basic approach (i.e., the customer-centric model) and incrementally migrating additional activities offshore over time. The desired end-state is achievable only over a period of time.
- Consider if there is potential for a long-term outsourcing arrangement at an offshore centre at the end of the delivery (i.e., a Design, Build, Execute arrangement). In such a case, the DC-centric model provides an additional advantage because there is no need for knowledge transfer to the customer personnel.

VIII. Model Refinement

The basic models are rarely applied on projects in their pure form. Instead, the engagement planner and managers usually refine the models based on specific aspects associated with their situations. The refinement process involves determining the best location for a given task.

For example, in the customer-centric model, the assembly or product tests can move from the customer site to the offshore centre site. Moving the assembly test to the offshore centre may be beneficial. Keeping all or a portion of the assembly test will result in removing more errors from the coded work units before they are transferred to the onshore centre or customer site.

Apply this fine-tuning process to all development tasks that lie on the border between the sites (e.g., application design). When considering moving a development task from its designated location in the model, consider risk mitigation strategies to address negative impacts of the move.

Apply the appropriate risk assessment criteria when deciding alternative locations for a given task. In general, the lower the aggregate risk results from looking across multiple risk factors, the more appropriate a given task is for execution at an offshore centre.

IX. Delivery Centre Organisation Structure

This work further explores the management structure, arrangement/contract, and staffing/organization required for completing the project successfully. The management and organisation structure has to be selected dependent on the distributed work model for successful and efficient development and delivery of the solution/ project.

There are two key aspects of the relationship between the customer and the delivery centre teams that set apart different Delivery Centre (DC) organisation approach:

- Extent of integration between the teams, i.e., the extent to which the DC personnel are engaged/used in the project's organization and the communication requirements between the customer site and the DC personnel.
- How much and which of the DC's methodology, processes, knowledge, tools, and technical facilities are used by the project?

The above two aspects influence the organisation structure of teams at customer and offshore centre sites:

- Communication: who is in control and communication and command lines

- Contractual/ Service Level Arrangements: what is sub-contracted and the arrangement details
- Recruitment/process: project staffing, process to follow

There are four different organisation structures which have been applied successfully in different areas. These approaches may not be applied in isolation as some practices are shared across and the boundaries are not rigid for successful completion of projects.

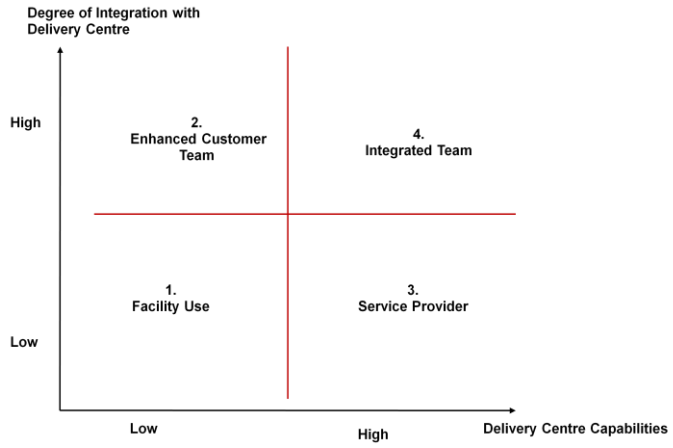


Fig. 7. Organisation structures

X. Facility Use

With this structure in place, the main aim of the customer is to get the office space from a DC on rent or lease along with the basic facilities such as desks, telephones, computers, shared file services, and connectivity to the customer site. This type of arrangement is more useful in the following circumstances:

- Projects which need to ramp-up quickly, which may be due to the following situations:
- A project having short time to market and the customer does not have time for building infrastructure or recruit staff quickly.
- Customers who do not have enough IT development space currently available in-house.
- Small projects which cannot afford higher cost of initial set-up in terms of both time and expense.

This organisation method has the following characteristics:

- Set up is easier and quick
- Influenced by the availability of infrastructure at the DC
- Capabilities of DCs are less used
- Little communication/ integration between DCs and customer
- This is not suitable when the distances between customer and offshore centre is large

XI. Enhanced Customer Team

This organisation method is used when the customer wants to enhance its capabilities by using the offshore resources in order to reduce the skills gap.

This organisation method has the following characteristics:

- Set up is easier and quick
- This is not suitable when the distances between customer and offshore centre is large (e.g., US with Philipines etc.)
- Capabilities of DCs are less used
- Higher level of communication/ integration between customer and DC
- Projects that will need a substantial number of skilled personnel at customer site and can be hired from DCs.
- The facilities and tools are not available at DCs.

XII. Service Provider

This organisation structure is can be defined as two teams, the customer site team and the DC team, working together in a highly collaborative manner. The work is subcontracted to the DCs with set of service level agreements and is managed with an established communication process. This type of arrangement is more useful in the following circumstances:

- A task is subcontracted to the DC, and the communication between the customer and DC is managed through a set of well-defined entry/ exit criterion.
- Generally projects/ customer will adopt fixed-fee costing method for this kind of structure.
- There is more freedom to DCs to allocate and manage the resources.
- The communication process will change as per the complexity of the project.
- This method requires clear defining of the accountability and delivery parameters.
- Demand process should be clearly defined in order to use the resources in an optimised manner.
- The customer will have relatively low start-up cost of contracting with the DC team as the DCs are using already existing methodology, training, and infrastructure.

Typical Use

- Projects that want to maximize the leverage of the DC capabilities and can work within the constraints of proven offerings, a known environment, and stable architectures. For example, an ERP project based on a well-understood offering (e.g., Oracle) with a stable platform, where a set of modules needs to be coded.
- Projects that anticipate their needs may expand rapidly in the future and need a choice of DCs that can

accommodate their requirements. For example, consider an SAP engagement in which demand rapidly surged, and it had contracts with three DCs to satisfy its capacity needs.

This organisation method has the following characteristics:

- Works well on projects with established application/technical architecture and with well-established and documented standards for specifying design deliverables (commodity market)
- Works well with offshore DCs
- Works well with established, mature offerings
- Works well in a fixed-fee arrangement to reduce risk of the customer site team
- Relatively low start-up costs
- Light-to-medium interaction between the customer site and DC teams
- High in leveraging the DC capabilities as the DC optimizes the use of its team and other resources
- To mitigate risks, mixing customer site and DC personnel is necessary and site liaisons could be a good option.

XIII. Integrated Team Approach

The project structure is similar to that of a general commitment, except that the project team is distributed among multiple sites. The project achieves significant cost savings by:

- Utilising the DC procedures, processes, tools, and infrastructure.
- Using the DC skills and resources by filling key technical and managerial roles with the DC personnel, and by integrating a critical mass of the DC personnel.
- Setting up accountability, which is less of an issue in this approach since the project is managed as one integrated team.

Typical Use

The project may have customer-facing or functional skills, but:

- It needs the DC for technical delivery capacity/expertise, e.g., the customer site team sold the work, but needs to assemble a team quickly to deliver the technical piece.
- The project relies on the DC to provide a significant portion or all of the technical delivery methodology, approach, estimating, etc.
- The project relies on the DC to fill some management and/or team lead spots.
- The project team in the DC operates as a virtual extension of the customer site team, with a fully mixed and integrated team of the customer team and the DC personnel.

Characteristics

- The DC team tends to drive many of the technical and project management approaches.
- This approach tends to have higher set-up costs.
- This approach supports projects that are in-flight, i.e., projects that started as traditional single-site projects and then become a multi-site project working with a DC. This is because this teaming approach accommodates knowledge transfer, which is part of the transition to multi-site process.
- It has the highest degree of integration between the customer and DC personnel.
- It achieves the highest leverage of the DC capabilities. It overcomes the Service Provider approach limitation by working well for projects that are based on new and existing application/technical architectures.

XIV. Choosing a Teaming Approach

Choosing the teaming approach involves assessing engagement requirements and examining a variety of factors. Some of these factors are listed below:

TABLE I. Teaming Approach

Issue/Factor	Facility Use	Augmnt. Customer Team	Service Provider	Integrat. Team	Comment
Quick/ fast Scalability	Suitable	Suitable	Depends on availability	Not suitable	The Integrated Team approach does not work for fast scalability because of the set-up costs/effort.
New Architecture	Questionable	Suitable	Not suitable	Suitable	Since a new architecture requires a high degree of interaction with the customer and the customer team, hence Service Provider type of interaction is not suitable.
Work-in-progress projects	Suitable	Suitable	Questionable	Suitable	Work-in-progress projects require knowledge transfer, so Service Provider approach is not appropriate.

Issue/Factor	Facility Use	Augmnt. Customer Team	Service Provider	Integrat. Team	Comment
Short time-to-market	Suitable	Suitable	Suitable	Suitable	If the project length is less e.g. 3 months or so, and an offshore centre is to be involved, then Integrated Team approach may not be suitable.
More knowledge transfer	N/A	Suitable	Not Suitable	Suitable	If higher knowledge transfer is required, then service provider approach is not suitable.

XV. Processes and Procedures for Onshore/ Offshore/ Near-Shore Work

The distributed onshore/offshore/ near-shore work arrangements require a number of steps to be completed. These are very much different from the traditional project management at one site. Therefore, organisations need to create a set of processes, procedures, tools, and techniques so that the distributed work can be managed effectively and efficiently. This helps organisations to manage and share the work across locations with a standard set of rules and processes. This ensures consistency and reusability of the resources/ documents and deliverables across projects.

Organisations can also get certifications like CMMI/ Six-Sigma or any other standard methodology for their processes, procedures, tools, and techniques. This is highly important to build confidence of customers in the delivery of projects on time and on budget.

XVI. Managing Transitions Across Phases/ Multiple Sites

Transition of project across multiple sites requires different set of processes, procedures, tools and techniques. The traditional transition processes of moving from analysis to design to build stages etc. may not be fully applicable. Therefore, organisation needs to define its new set of processes to manage the work/ project effectively.

Organisation must consider the following issues to prepare the plan and manage the work:

- What is the best possible and optimum way of transfer of knowledge from one site to another
- Monitoring and Controlling process
- Dry run of the project
- Managing risks
- Approval of the transition process by the stakeholders
- Prepare checklists for various stages of the project
- Prepare contingency plan
- Skills-gap analysis for resources

Keeping in mind the above issues, the following are some of the effective techniques to manage the transfer of the project to the development centre.

- Process for Knowledge transfer: Organisation must use a good process to transfer the knowledge from one site to another and this has to be measured against baselines to make it efficient.
- Managing with Checklists: Checklists are created for various modules, deliverables, documents, hardware, software, databases, resources and skills. These are very effective in controlling and seeing the progress of the project.
- Dry run/ pilot run: Project is given a dry run for a few set of data to see that the overall objectives are met and project is behaving as per the expectations before the final release and go-live.
- Reduce communication gap: Do regular secure information sharing with stakeholders. During transition, see the possibility of having key users can work at development centre.

XVII. Configuration Management (CM)

In the multi-site environment, the most affected area is configuration control. Organisation must create set of processes, procedures, tools and techniques to manage the integrity of the project across sites and ensure various stages are completed as per the plan.

The repository for the configuration management must be able to provide service to all the sites with ease and flexibility and also adhering to the various security concerns.

Organisation has to consider following questions:

- Has the CM plan/approach been defined?
- Has the change process been defined and approved?
- Has the CM effort been estimated and budgeted?
- Is there involvement of resources from the delivery centre?
- Have you identified roles for CM support activities?
- Have you signed an agreement/SLA with the delivery centre for CM support?
- Has ownership for all files/objects been assigned?
- Has long-term ownership of the CM repository been resolved?
- Have contingency and roll-back plans been established in case the repositories get out of synch?

- Have plans for CM audits been addressed in the CM plan, and are they covered by the CM budget?
- Are there plans to test the CM repository from all remote locations to ensure that accessibility and performance requirements are met?
- Approval of CM plan by stakeholders.
- Connectivity issues: The connection speed, bandwidth, and cost influence where the repository can be located and which CM tools to be used. Various options are VPN, Leased line, Company-WAN, etc.
- Where would the repository be located?

Three approaches for organizing a CM repository are identified as best practices: centralized, independent repositories, and multi-site with replication. These approaches differ from each other in terms of performance, flexibility, and cost.

Centralised Repository: This offers high flexibility, easy set-up and operate, easier regulation and compliance due to single site but its performance is dependent on the connectivity.

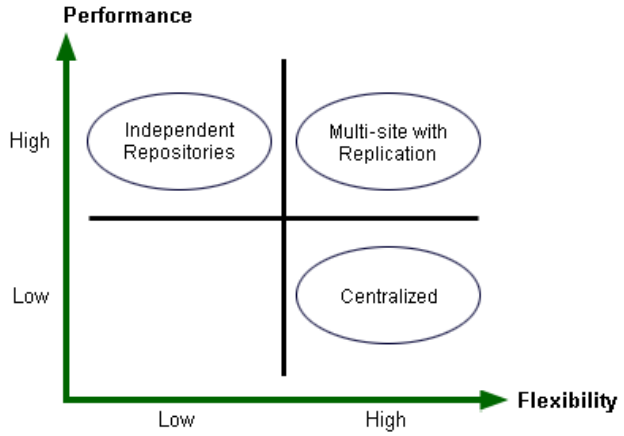


Fig. 8. Potential Repository Locations

Independent repositories: Low Flexibility, medium set-up complexity, high performance as individual site has its own CM tools, files are synchronised manually.

Multi-site with replication: High performance, Higher flexibility, but with higher set-up costs, requiring high-speed and bandwidth connection

- What are the various CM Tools?
- Availability
- Number of sites
- Project Size and complexity

XVIII. Estimation Process

Estimation process is difficult for the distributed work, as the number of parameters is more due to the involvement of multi sites. Significant effort has to be put into arriving at the estimate of time and cost for completion of the project. Development centre must be involved in the estimation process in order to minimise the risk. Organisation must take into account the followings issues:

- Has the estimation for time, effort and cost been done and approved by appropriate stakeholders?
- Have you created service level agreement and approved by stakeholders?
- Have you planned for time and budget for training of offshore resources?
- What is the plan for knowledge transfer and budget as well as time frame for the same?
- Have the estimates considered risk factors such as lack of communication, cultural issues, resource availability, and technology differences, etc., which are common in multi-site development? What kind of buffer is available?
- What is the contingency plan?
- Are all key areas covered in the estimates: analysis, design, build, test, etc.?
- Is the cost for monitoring and controlling also be estimated?
- Have you involved all stakeholders, technical and functional, to assist the estimating?
- Have you allocated budgets across organizations/locations and assigned responsibilities for deliverables?
- What are the expenses for travel, communication between sites, etc.?
- For costing the project, have you involved the delivery centre experts in providing rate, tax, multi-year inflation adjustments, etc. into the cost calculations?
- Have you considered any pre-existing master services agreement conditions that you may already have with the customer in terms of pricing this new deal?
- Have you accounted for currency and inflation risks (expenses will be through local currency)?

XIX. Intercultural Guidelines for Distributed Work

It is crucial for today's business personnel to understand the impact of cross cultural differences on business, trade and internal company organisation. The success or failure of a company, venture, merger or acquisition is essentially in the hands of people. If these people are not cross culturally aware then misunderstandings, offence and a breakdown in communication can occur.

The need for greater cross cultural awareness is heightened in our global economies. Cross cultural differences in matters such as language, etiquette, non-verbal communication, norms and values can, do and will lead to cross cultural blunders.

U.S. and British negotiators found themselves at a standstill when the American company proposed that they "table" particular key points. In the U.S. "Tabling a motion" means to not discuss it, while the same phrase in Great Britain means to "bring it to the table for discussion."

Cultural awareness is crucial for any development project involving multiple countries or workforces. Differences in culture can affect team communication and influence team processes. This has always been an aspect of project work, and will become increasingly prevalent as more and more projects use multiple development sites and local and global workforces. It is important to value the diversity of people and practices across the world. The company's underlying code of ethics and positive support of people through company-wide programs are key pillars of running any successful engagement. Organisations should consider the following issues for effective communications across different countries and cultures:

- Increasing cultural awareness
- Identifying a communication strategy to overcome language barriers
- Encouraging team work
- Providing opportunities for face-to-face interactions
- Using effective virtual teaming tools
- Addressing country-specific business hours and holidays
- Groups vs. individual orientation
- Hierarchy and status
- Risk taking ability
- Communication Style – Direct/ Indirect
- Task vs. relationship
- Short term vs. long term
- Use of implicit and explicit messages
- Tolerance for ambiguity
- Responses to problems
- Use of silence for showing respect vs. asking questions up-front
- The desire to please others vs. the desire to identify issues.
- The desire to preserve other people's dignity and self-respect.
- Different emphasis on time.
- The desire for perfection.
- A strong social network.
- A strong work ethic.

XX. Common Expectations

Language skills are a key part of working across geographies, and English is often the most common business language used. Accents may initially cause a few issues.

If there is a language barrier, identify a communication strategy to overcome it.

- Identify leaders with good language skills as key contacts and include them on all project status calls.
- Some people have good language skills, but may not be as confident as others. Some, who may feel less comfortable in the multicultural work environment, are likely to be more timid in discussions. During meetings, explicitly invite them to speak their thoughts and opinions.

In general, multi-site projects use extensive written communications to minimize misunderstanding verbal messages. Instant messaging tools can be an effective substitute for telephone conversations in circumstances like this.

Some cultures are not accustomed to writing in English at the volume that projects require, so use a combination of written and verbal communication that makes sense to the overall project team.

Organisation should consider the following questions for improving cross cultural awareness:

- Be aware of your own culture. What is your communication, decision making, and issue management style?
- Did you learn about the culture of global colleagues?
- What are the plans to raise cultural awareness across the project team?
- Are you aware of the potential cultural differences that affect your project's communication, decision making, and issue management?
- How will you respond to these cultural differences as a project and as an individual?
- What plans do you have to promote collaboration and communication?
- Have you communicated these plans to both the customer (local) site and the global teams?
- Have you trained both the customer site and global teams to use the virtual teaming tools effectively?
- Have you met the teams from the different geographies?
- How will you measure that your multi-country and/or multi-workforce project team is communicating effectively?

XXI. Issue/ Problem Management

Issue/problem management involves the process for identification, analysis, resolution, reporting, and escalation of the project's issues and problems. There has to be clear documentation of how and with which parameters issues are prioritized, assigned, communicated, viewed, escalated, and resolved.

With multiple sites and lesser face to face communication, resolving issues and problems is more difficult. Therefore, teams at different sites will have to rely on a common process and/or an automated tool to track, share, and resolve issues/problems in a timely manner.

Organisation should consider the following parameters for managing the issues effectively and efficiently:

- Plan issue/problem management.
  - Define the issue/problem management objectives and goals.
  - Define the issue/problem management process. Include escalation procedures.
  - Identify issue/problem management roles.
  - Identify issue/problem management tools.
  - Finalize issue/problem management plan. Ensure all sites understand and agree to the plan.
- Execute issue/problem management processes.
  - Identify issues/problems.
  - Track issues/problems.
  - Assess issues/problems.
  - Develop issue/problem resolution.
  - Monitor and communicate on issues/problems.
  - Report metrics.

Organisation should consider the following questions:

- Is an issue and problem management process established?
- Have you selected issue/problem management tool(s)? What is the installation/roll-out plan for the tool(s)?
- Are issue/problem management roles defined and assigned?
- Are issue/problem documentation standards defined?
- Has an escalation process been established?
- Have you developed a plan for communicating issues/problems to team members and the customer?
- Does training exist for those who use the issue/problem management tool(s) and processes?
- Were metrics created to measure the effectiveness of the issue/problem management process?
- Have you done a causal analysis of the issues at defined milestones?

XXII. Resource Management and Organisation Design

Understanding Organisation design is very important so that various challenges of current capability assessment, enterprise environmental factors like work, culture, management style, etc. can be addressed for organising a distributed project team, define project roles, and manage the resources. This will help stakeholders estimate the work effort, and plan for the work, and efficient use of resources and communicate clearly the roles and responsibilities. In order to manage the distributed effectively and efficiently, stakeholders form all areas must be involved in planning, and build team behaviour and not Offshore vs. onshore team /client team.

Approach

Organisation structure and design could be as follows:

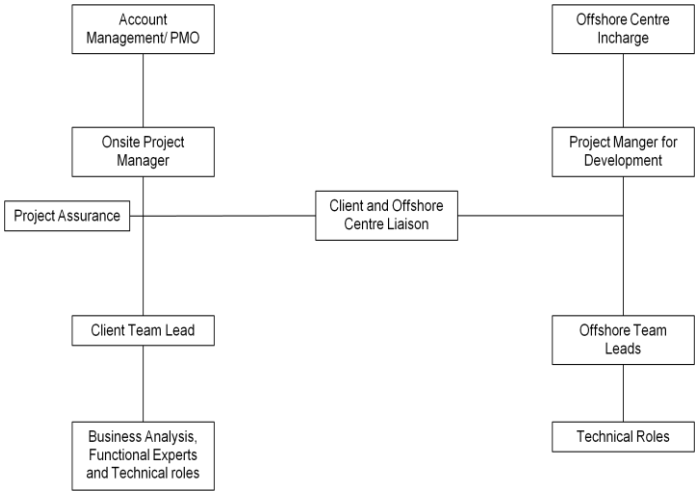


Fig. 9. Standard Organization Structure

- Account Management/ Program Management Office takes care of the all the projects being developed with customer. This is required to care of business requirements and customer stakeholder expectation management.
- Offshore Centre In-charge is responsible for all the projects running at the centre.
- Onsite Project manager takes care of the full project, managing schedules, budgets, project execution, monitoring and controlling. Project Manager communicates all the project progress to all the stakeholders directly and through Client and Offshore Centre Liaison. Project Assurance manages and communicates the status of the project to all stakeholders and monitors the project risk and escalates the risk/ issues as required.
- Offshore Centre manager for Development is responsible for managing the work at offshore centre and also provide status report to the onsite project manager.

- Client and Offshore Centre Liaison is to improve the communication between the onsite and offshore centre and provide information to all stakeholders to minimise the risk.
- Team leads at individual sites manages their respective teams for performing various business, functional and technical roles to complete the project as per the plan and manage the resources efficiently and effectively.

In addition to making organization and staffing of the project more complex, multiple sites also makes managing the resources more complicated than with traditional, one-site projects.

Project manager and stakeholders must take care of the following:

- Leverage delivery centre resources as much as possible when staffing projects to take advantage of the deep application and technical skills and cost savings. Engage delivery centres early to secure resources.
- Subject matter experts (SMEs) are needed from the DC to help define and refine the estimates and work plans.
- Each centre is different from other in the staffing model, resource management, processes and procedures. Work closely with delivery centre liaisons so that right skills people can be identified quickly. This will help you avoid delays in obtaining resources. Understand the delivery centre's demand management processes so that necessary lead times can be accounted for in the project schedule.
- The cost structure associated with delivery centre resources varies.
- Liaisons can help you to find resources and guide through the complexities of identifying and procuring offshore resources in a manner that complies with company and national labour policies – for example, visa, wages and expenses, and tax considerations.
- UK work permit process is different from the USA and also the time required is different. Visa lead times vary by country of application (India vs. China) and by cities within a country (Bangalore vs. Mumbai). Visa lead times also vary by visa type (H, L, etc). The lead-times vary over time as per new government legislations from time to time.
- Discuss about the management style: dual management or not; long-term planning for resources, fully utilising the resources from offshore centre, etc.
- Is the project to be released in multiple stages? What are the plans for multiple releases?
- How will you take care of attrition of skilled resources?
- Roles and responsibilities are clearly defined for each team member including owner, reviewer, and approver of the various deliverables and milestones in each stage. Consider bringing offshore resources onshore and vice versa for better understanding and also transition of work/ tasks.
- To break cultural barriers, involve people from different teams and form virtual sub groups.
- Treat each member equally even though their parent organisation policies may be different e.g. vacation and holidays, working hours, overtime, and flexible work hour policies.

The following points may be considered for an effective organisation design and resource management in distributed environment:

- Have all stakeholders (customer site, delivery centres, users, 3rd party vendors, etc.) been considered, when defining the organization design and resource needs?
- Consider involving client for organisation design and resourcing needs.
- Early notification to the delivery centre resourcing personnel during the selling process to tell them that a deal is under progress in which they may be involved?
- Involvement of subject matter experts from all sides for proposal, estimating, and planning of the project.
- Has an organization design and hierarchical structure been defined and approved by all the stakeholders?
- Is there roles and responsibility document and matrix? Does everyone agrees and approves it?
- How will the third parties be integrated? What are their roles and responsibilities and deliverables?
- What are the communication processes and requirements for the current project?
- Is the offshore centre being used for only for application and technical skills? What will be the cost savings?
- Does the project need contract staff for filling in the skills gap?
- Does the project management overlaps with other projects?
- From the project requirements and scope, pan effectively and efficiently for the future demand of resources.
- Consider the appraisal process and career progression path of the offshore resources. This should be managed as per the needs of the centre.

XXIII. CASE STUDY

The purpose of the group ERP Consolidation project (GERP) is to implement one SAP based ERP application that will support standardised and simplified business processes for all of the group businesses in organisation-Z. The project planning and management was done as per the PMBOK process and knowledge management areas.

This project is classified as a Business Initiative. Benefits will arise from the lower cost of ownership of a single consolidated ERP system for the Service Companies and the reduced cost of support through offshoring a significant part of the new support organization required to support the consolidated application. Additionally, business benefits will arise through the consolidation of back office functions enabled through use of simplified, standardised business processes and systems.

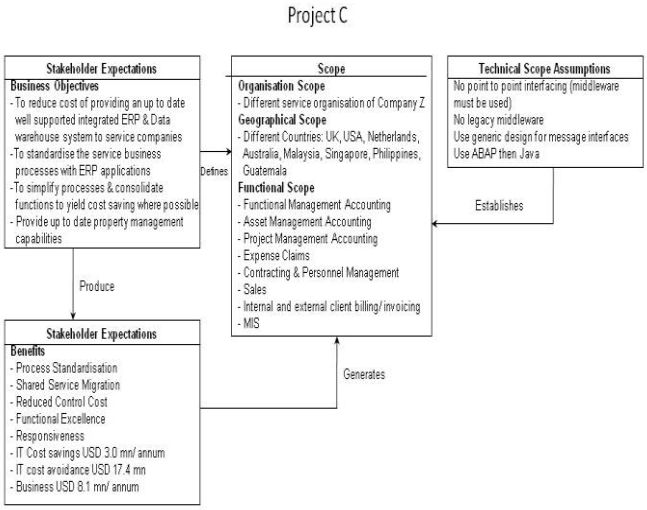


Fig. 10. Project 'C'

The main business objectives of this project are:

- To reduce the cost of providing an up to date, well supported integrated ERP and data warehousing system to the Companies.
- To standardize the business processes with ERP applications.
- To simplify Services business processes and to consolidate functions where possible to yield cost savings in operating those functions.
- Provide up-to-date Property Management capabilities where organisation's Real Estate Services can consolidate property information and can standardise business processes associated the administration of organisation's property.

The project objectives are in line with the CMD, CFO and other who endorse Group ERP Strategy based on SAP software. The project objectives are also in line with IT to reduce IT application support costs across the Group through rationalization of IT applications and offshoring of application support. The project objectives are also consistent with the recommendations to use consistent processes across the group and supported by one common system. Additionally, the project objectives are also consistent with the finance strategy to standardize and simplify financial processes, provide increased transparency of financial information and a consistent controls framework.

XXIV. IT Operational Cost / Benefit

To estimate future IT operational cost, the Operational Cost information for the individual Service applications was collected and decomposed into three areas; ERP Cost, Data Warehouse Cost, and Other ERP Related Costs. Each of these cost components, was further broken down by: Hardware, Software License Fees, Application Support and Run & Maintain Enhancements Costs.

From this base information, collected from the focal points, forecasts were made using knowledgeable resources, accepted estimating models and assumptions based upon best information.

An estimated $3.1 million in benefits may be obtained in IT operational cost by consolidating the Services businesses on to one ERP application.

The following are the guidelines followed to determine the portion of the project costs that should be considered capital and expense.

- Program & Management costs
  - Strategic investments required to deliver the system
  - 60% capital and 40% expense
- Implementation costs
  - Development predominant activity
  - 100% capital
- Training and data conversion
  - 100% expense
- Post go-live operational & support
  - 100% expense
- Post go-live upgrades
  - 100% capital

XXV. Recovery Mechanism

Ownership of the GERP project is based upon a cost recovery model where all participants share in the ownership of the intangible asset. The premises for the ownership and cost allocation is:

- Single entity captures costs associated with GERP
- Periodically (quarterly) cost are passed to the participating entities
- Capital cost are recorded as work in progress
- At go-live benefiting business entity reimburses and records intangible asset and amortize asset over 5 years
- Recommend payment based on named number of users and any unique customization charged to requesting entity
- Payment "trued-up" upon completion of project

XXVI. STRATEGIC / INTANGIBLE BENEFITS

Additional strategic and intangible benefits associated with the consolidation of the Services ERP and data warehouse applications have been identified (but these are difficult to quantify). Benefits include:

- Faster and less costly implementation of new strategic initiatives
- Platform available for any future new Business Service or Functions inclusion of which should lower costs for all participants
- Easier sharing of best practices
- Facilitates off-shoring/outsourcing.
- Facilitates improved controls and compliance
- Common processes and formats for customers
- More flexible workforce
- Enhanced decision making through more readily available and higher quality Management Information
- Easier benchmarking across Business Services and Functions
- Consolidated view of services position across customers/suppliers.

XXVII. PROJECT MANAGEMENT

The project planning and management was done as per the PMBOK process and knowledge management areas. Various documentation and deliverables were created along with milestones. The project was managed using multi-centre scenario as given below:

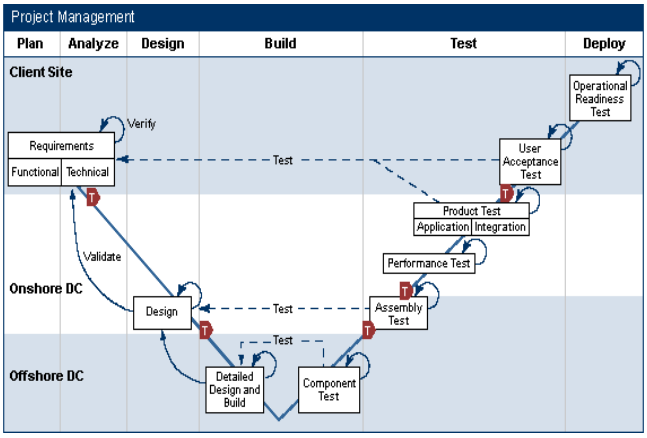


Fig. 11. Multi Centre Scenario

All the requirements were gathered at all the locations and design was validated at the onshore site and detailed design done at offshore DC along with various component tests and part of assembly tests. Solution was then implemented at the onshore sites in various countries and final testing at client locations.

This model provided the benefits of the both the DC-centric and customer-centric models. Cost-savings achieved by using the offshore centre and the risk was reduced because the customer team works closely with onshore/near-shore centre.

XXVIII. PROJECT TIMELINE

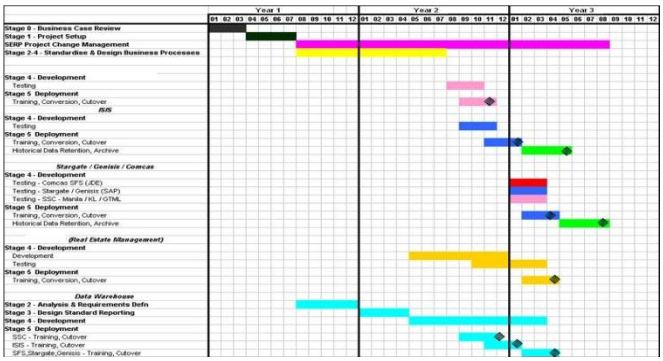


Fig. 12. Timeline

The Services ERP project will implement a rigorous risk management process, which will identify potential risks, qualify their probability of occurrence, quantify their potential cost and time impact, and define risk mitigation and avoidance strategies.

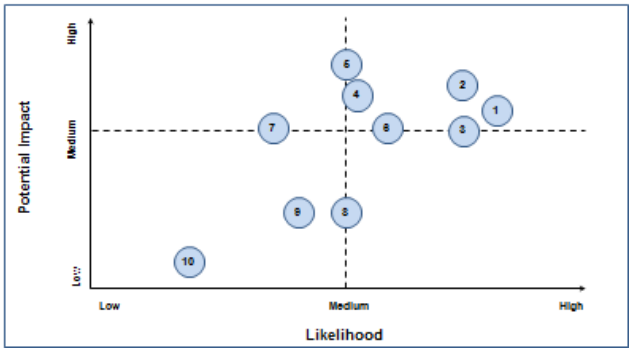


Fig. 13. Risk Prioritisation

Above diagram shows the risks are identified as 1.Internal resources, 2. Organisational Change Resistance, 3.Historical Data Retention Requirements, 4. Group Consultation/ Documentation, 5.Benefits Realization, 6. Group Business Process Requirements conflicts 7. Delay with group ERP implementation 8. Business Reorganisation 9.Project Cost overrun, 10. Higher infrastructure costs.

Risk response planning was also created and monitored and controlled by the project management office.

XXIX. VALUE PROPOSITION AND ECONOMICS

Consolidation of these four different ERP systems and data warehouse systems delivered approximately 25% reduction in IT related cost, as well as potential business cost savings enabled by the consolidation of these systems and standardisation of business processes.

Summary of Savings

Business	$ 8.1m p.a.
IT	$ 3.4m p.a.
IT Cost Avoidance	$ 17.4m

The benefits are described in more detail below.

XXX. Business Cost Savings

Implementing and centralising standard processes generated approximately USD 7.4 m p.a. of the USD 8.1m cost savings. Financial closing and central master data maintenance are examples of where cost savings can be realised. This resulted in reduction of business resources to support GERP and allowed moving more operational functionality to low-cost centers. Further cost reductions were realised for a reduced number of annual system audits (.2m p.a.) and .5m in simpler implementation of future Group initiatives like SOX documentation, International Accounting Standards adoption, Global Credit Card rollout, etc.

The GERP Application is a key enabler to significant changes in the management of Services business processes and related financial information. One such potential change is the simplification and standardisation of the intra and inter-company billing process. Standard intra and inter-company process on a single ERP platform will facilitate additional efficiencies in the Shared service centers. Central HR benefited from improved intra and inter-company processes through reduced number of interfaces of payroll information from employees and reduced number of applications that require account analysis. Reduced invoice volumes, standardised customer invoices, improved data integrity and fewer resources doing internal business will result in additional efficiencies. Benefits were also achieved from consolidation of master file data maintenance and financial closing functions into a common back office.

XXXI. IT Cost Savings

USD 3.4m annual savings in IT operating cost are estimated through reduction of ERP and data warehouse applications to one consolidated system. Reduced application support costs drive the largest savings in IT cost from approximately USD 5.3m to USD 2.5m. This USD 2.8m saving is due to the reduction in the number of FTE's required to support the application and off shoring of application support as per ICT Vision. The overall system enhancements expenditure reduced somewhat through avoidance of duplicated spend.

The savings in system enhancements is USD 0.5m per annum. Real Estate Services realised approximately USD 0.3m p.a. savings by replacing the ABC application with the Property Management functionality transferring to SAP and other functionality to other standard packages.

XXXII. IT Cost Avoidance Benefits

A total of approximately USD 17.4m has been identified in one-time cost avoidance benefits. This is comprised of a USD 4.5m required upgrade of XYZ in earlier to a supported SAP version. The current XYZ SAP version (x.x) is supported through a temporary arrangement with annual cost increases and will become increasingly difficult to support and adapt to business needs.

Without one standard ERP, inconsistent financial processes and controls across the Services and Functions would have remained and above benefits would not have been realised. In addition there was a continued risk of failing to achieve lower cost finance function without GERP.

XXXIII. Conclusions

With increasing globalisation, organizations are now using more and more distributed work environments and the management of such large distributed projects is always complex and difficult. This paper discussed various models, processes and flows for the effective and efficient management of distributed or onshore/ offshore projects. Four key models were described along with their characteristics, their advantages/ disadvantages and the best possible scenario in which each is applicable. It also focused on the teaming and organization structure approaches. Various advantages and disadvantages of each teaming approach were also discussed along with the selection criterion for project/ situation.

Earlier research focused on discussing very simple techniques/ processes and very basic organisation structure, but could not clearly define models how the work will be distributed among onshore, near shore, and offshore centres. In our research, four models and four teaming approaches are discussed, highlighting the importance and selection criterion, characteristics, and their best scenarios for use.

A case study of one of the projects using one of the models (i.e Multi Centre Scenario) has shown that major benefits could be achieved. These benefits are highlighted as business cost savings, IT cost savings, and IT Cost avoidance benefits. The project planning and management was better and the project was delivered on time with improved and enhanced project monitor and control mechanism.

All other project management knowledge and process areas of PMBOK were used effectively and efficiently. All the documents, deliverables were created as per PMBOK and milestones monitored and controlled to deliver project in various countries.

A very large number of organisations now manage projects globally and use some kind of process for managing projects in different countries. The models and teaming approaches defined here will be highly beneficial to such organisations as this paper describes a better structured flow, processes and organisation structure to manage global/ distributed projects effectively and efficiently.

References

[1] Armstrong, D.; Cole, P., "Managing Distances And Differences In Geographically Distributed Work Groups" in P. Hinds & S. Kiesler (ed.) Distributed work, MIT Press, 2002, pp. 167-186

[2] Salger, F.; Englert, J.; Engels, G., "Towards Specification Patterns for Global Software Development Projects - Experiences from the Industry", 7th International Conference on the Quality of Information and Communications Technology (QUATIC), Portugal, 2010 , pp 73-78

[3] Salger, F.; Sauer, S.; Engels, G.; Baumann, A., "Knowledge Transfer in Global Software Development - Leveraging Ontologies, Tools and Assessments", 5th IEEE International Conference Global Software Engineering (ICGSE), USA, 2010, pp 336-341

[4] Narayanan, Sidharth; Mazumder, Sumonta; R., Raju, "Success of Offshore Relationships: Engineering Team Structures", International Conference on Global Software Engineering, ICGSE'06, USA, 2006, pp 73- 82

[5] Persson, J.S.; Mathiassen, L.; Boeg, J.; Madsen, T.S.; Steinson, F., "Managing Risks in Distributed Software Projects: An Integrative Framework", IEEE Transactions on Engineering Management, vol. 56, Issue: 3, 2009 ,pp 508–532

[6] Khan, H.H.; Malik, N.; Usman, M.; Ikram, N., "Impact Of Changing Communication Media On Conflict Resolution In Distributed Software Development Projects", 5th Malaysian Conference in Software Engineering (MySEC), Malaysia, 2011, pp 189-194

[7] Lane, M.T.; Agerfalk, P.J., "Experiences in Global Software Development - A Framework-Based Analysis of Distributed Product Development Projects", 4th IEEE International Conference on Global Software Engineering, ICGSE, Ireland, 2009, pp 244 – 248

[8] Niinimaki, T., "Face-to-Face, Email and Instant Messaging in Distributed Agile Software Development Project", 6th IEEE International Conference on Global Software Engineering Workshop (ICGSEW), Finland, 2011, pp 78 - 84

[9] Czekster, R.M.; Fernandes, P.; Sales, A.; Webber, T., "Analytical Modeling of Software Development Teams in Globally Distributed Projects", 5th IEEE International Conference on Global Software Engineering (ICGSE), Ireland, 2010, pp 287–296

[10] Bartholomew, R., "Globally Distributed Software Development Using An Immersive Virtual Environment", IEEE International Conference on Electro/Information Technology, EIT, USA, 2008, pp 355-360

[11] Hashmi, J.; Ehsan, N.; Mirza, E.; Ishaque, A.; Akhtar, A., "Comparative Analysis Of Teams' Growth In Offshore And Onshore Software Development Projects", IEEE International Conference on Management of Innovation and Technology (ICMIT), Singapore, 2010, pp 1163–1167

[12] Hinds, P.J.; Bailey, D.E., "Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams", Organization Science, 2003, vol. 14 (6), pp 615-632

[13] Swan, Bret; Belanger, France; Beth Watson-Manheim, Mary, "Theoretical Foundations for Distributed Work: Multilevel, Incentive Theories to Address Current Dilemmas", IEEE 37th Hawaii International Conference on System Sciences, Hawaii, vol. 1/04, 2004, pp 1-10.

[14] Bailey, D. E.; Kurland, N. B., "A Review Of Telework Research: Findings, New Directions, And Lessons For The Study Of Modern Work", Journal of Organizational Behavior, vol.23, 2002, pp 383-400

[15] Pinsonneault, A.; Boisvert, M., "The Impacts of Telecommuting on Organizations and Individuals: A Review of the Literature", in "Telecommuting and Virtual Offices: Issues and Opportunities", Johnson, N.J., London: Idea Group Publishing, 2001, pp 163-185

[16] Pearlson, K.E.; Saunders, C.S., "There's No Place Like Home: Managing Telecommuting Paradoxes", Academy of Management Executive, vol. 15, 2001, pp 117-128

[17] Belanger, France; Beth Watson-Manheim, Mary; Jordan, D.H., "Aligning IS Research and Practice: A Research Agenda for Virtual Work," Information Resources Management Journal, vol. 15, 2002, pp. 48-70

[18] Igbaria, M., "The Driving Forces in the Virtual Society," Communications of the ACM, vol. 42, 1999, pp 64-70

[19] Alveson, M, "Knowledge Work and Knowledge-Intensive Firms". Oxford University Press, New York, 2004

[20] Hornett, A., "The Impact of External Factors on Virtual Teams: Comparative Cases", in Pauleen, J. (ed.), "Virtual Teams: Projects, Protocols and Processes", Idea Group Publishing, UK, 2004

[21] Turkington, D., "Remote Resourcing", The Beca Infrastructure Board, Auckland 2004

[22] Bélanger, F.; Collins, R.W., "Distributed Work Arrangements: A Research Framework", The Information Society, vol 14, 1998, pp 137-152

[23] Cramton, C.D., "Attribution in Distributed Work Groups", in Hinds, P.J.; Kiesler, S. (ed.) "Distributed Work", MIT Press. London, England, 2002, pp 191-212

[24] Mohammad Jafari, M.; Ahmed, S.; Dawal, S.Z.M.; Zayandehroodi, H., "The Importance Of E-Collaboration In SMES By Project Management Approach A Review", 2nd International Congress on Engineering Education (ICEED), Malaysia, 2010, pp 100–105

[25] A Guide to the Project Management Body of Knowledge, 5th edition, PMI, USA, 2013

System Autonomy Modeling During Early Concept Definition

Rosteslaw M. Husar
Southern Methodist University
Dallas, TX 75205, USA

Jerrell Stracener , PhD
Southern Methodist University
Dallas, TX 75205, USA

***Abstract*—The current rapid systems engineering design methods, such as AGILE, significantly reduce the development time. This results in the early availability of incremental capabilities, increases the importance of accelerating and effectively performing early concept trade studies. Current system autonomy assessment tools are level based and are used to provide the levels of autonomy attained during field trials. These tools have limited applicability in earlier design definition stages. An algorithmic system autonomy tool is needed to facilitate trade off studies, analyses of alternatives and concept of operations performed during those very early phases. We developed our contribution to such a tool and described it in this paper.**

Keywords—Systems Engineering; Autonomous Systems; Requirements Engineering; System of Systems component; System Autonomy Modeling

I. INTRODUCTION

The United States Department of Defense (USDoD) is facing declining defense budgets for at least the next several years while adversary nation experience double digit defense budget increases1. In this fiscal environment, the USDoD must find new ways in meeting the goal of providing national security. A significant portion of the budget is for manpower in the operations and support phase of the system life cycle. Unmanned autonomous systems can provide this force multiplier2 allowing a single operator to manage multiple unmanned systems[9][24]. Autonomy for unmanned systems is the needed technological innovation to reduce the workload of human operators. This technological demand is greatest in military operations where significant loss of life and extreme hazardous situations are common place.

Unmanned vehicles are a key component of the U. S. Navy (USN) defense transformation[28]. The USN has several programs under development to address reduced manning with increased use of unmanned vehicles (UxV) [32]. These unmanned vehicles require a significant amount of human interaction (HI) to control the UxV and interpret a significant amount of down linked data. Assessing intelligence, surveillance and reconnaissance (ISR) data to develop actionable security operations will continue to be a national priority. The amount of data collected is overwhelming the analysts. Current state of the art unmanned systems, like the Predator Unmanned Air Vehicle, require a sizeable team to operate the air vehicle, interpret sensory information, dynamically assess mission impacts and execute missions. The increasing demand for ISR missions are increasing crew support, counter to declining budget trends.

A. Background

The Congressional Budget Office in their FY2014 report anticipates that the portion of Gross Domestic Product (GDP) dedicated to the USDoD will continue to decrease over the next several decades[11]. Future reduced funding for systems development will take a larger share of the operations, maintenance and personnel costs within the constrained budget[96].

To address this environment of declining defense budgets concurrent with increasing threats, the U. S. Navy is implementing unmanned technology in meeting the goal of providing national security at reduced cost[4]. Autonomy is the needed technology to reduce manpower by allowing a single operator to manage multiple unmanned systems.

Autonomous systems results from complex integration of human intelligence and machine automation capable of adapting to unforeseen events[4]. Autonomous systems could operate more independently and with lower focus levels of human interaction (HI), thus allowing for significant reductions in manpower.

The USN has several programs under development to address reduced manning through increased reliance of unmanned vehicles (UxV) and these systems require ever increasing levels of complex automation and autonomous capabilities. Proposed near-term maritime missions involve the use of collaborative unmanned autonomous systems.

B. Information Technology Acquisition Changes

The 2009 & 2011 National Defense Authorization Acts, Sec 804, mandated a new Information Technology (IT) Acquisition Process, Fig. 1[30], was required because:

The Defense acquisition process structured for weapon systems was ill-suited for information technology and

- Systems take too long to deliver and inconsistent with technology cycles;
- Documentation intensive, time consuming and process bound to respond effectively to end-user needs;

[1] Karl Ritter, April 15, 2013, The World Post and the Stockholm International Peace Research Institute (SIPRI)'s Year Book 2013 summary on military expenditure reported defense budget increases for China of 325%, Russia of 179% and South Korea of 59%.

[2] A capability that, when added to and employed by a combat force, significantly increases the combat potential of that force and thus enhances the probability of successful mission accomplishment. http://www.military-dictionary.org/force_multiplier

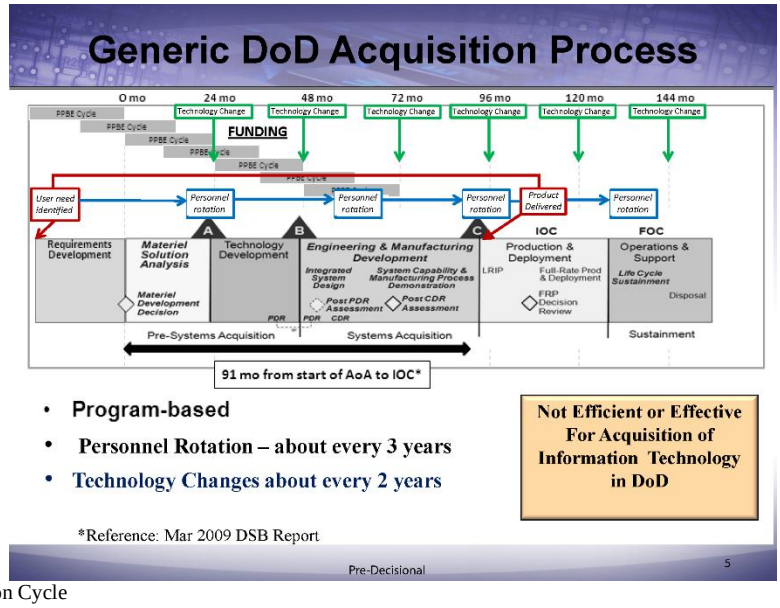


Fig. 1. Long System Acquisition Cycle

- Oversight process not aligned with rapid acquisitions (favors large programs, high-level oversight) and
- Lack of accountability by personnel in the oversight process;
- Complexity inherent in aligning three major USDoD processes - Requirements, Resourcing and Acquisition;
- Funding process inconsistent with pace of evolving mission requirements;
- Current metrics (financial, acquisition process) don't work well in measuring IT success;
- Lack of meaningful trades between performance, cost and date-to-field;
- Overly detailed requirements that are inconsistent with pace of technology change and need for rapid delivery;
- Inability to prioritize requirements effectively;
- Testing is integrated too late and serially;
- Cyber-security is inadequately managed during the acquisition process;
- Lack sufficient numbers of individuals with proven records of acquisition success;
- Significant cultural impediments to change.

What is common across these definitions is the need to develop the "best" end product in response to a set of needs. This can be accomplished by execution of the systems engineering process where the requirements analyses and the allocation of those requirements to the Functional Analysis and Allocation are performed [10]. Integrated tools to support analyses and assessments are critical at this early design phase because any shortfalls or miscalculations become costly if carried through the life cycle [1]. The AGILE methodology is such an engineering process and focuses more on the collaborative efforts between the software developers and 'customers to allow for early capability releases [2][21]. As a result, the releases are time driven rather than event driven which allows for maturing of the capabilities based on 'customer feedback'. This accelerated and iterative development release model is reliant on rational tools to support system analyses and requirements trade off studies as design deficits or errors become costly at later stages of the product life cycle. The AGILE methodology is appropriate for capabilities realized by software rather than implemented by hardware, which requires longer procurement and fabrication cycles [11] [33].

II. System Autonomy Assessment

The USN has defined that autonomous systems results in a complex integration of human intelligence and machine automation capable of adapting to unplanned events changes encountered during mission operations [17]. Current models assess system autonomy by assigning single numeric levels and do not support requirements trade off studies [4]. On the low end of the autonomy scale, (tele-operations), a computer offers no assistance and the human operator must take make all decisions and actions[28].

Complete autonomy is at the other end of the scale as a computer decides everything and ignores the human being. As systems become more complex and the need for collaboration between these subsystems increases, a single numeric level describing autonomous capabilities is not adequate. Missions are becoming more complex and require systems of autonomous systems architectures that dynamically adapt to the varying levels of autonomous operations needed. Understanding the complex and dynamic relationship between human interaction, machine autonomy and the mission operational environment is critical in early candidate architecture trade studies.

AoA of system architecture designs have a significant impact on the mission concept of operations (CONOPS) and

must be efficiently done in those early stages of development[33]. An approach to characterize autonomy in the early requirements modeling and trade-off studies is critical as large systems of systems development efforts are now quite commonplace [3]. Although a significant body of work exists to assess autonomy, a mathematical relationship, as addressed by the Defense Science Board, does not exist to study the impacts of reduced manning and machine automation to meet mission success. Current methodologies and frameworks used have led to a misunderstanding of the level of autonomy required and developed.

While often interchanged, 'automation' and 'autonomy' are not synonymous and what is frequently referred to as a level of autonomy 'is a combination of human interaction and machine automation' (USN Chief of Naval Operations). The CNO continues to state that 'the degree of machine automation is not easily categorized' and not fully 'understanding autonomy has hindered development' of unmanned systems in the Navy.

As the USDoD acquisitions favor decreasing and rapid development cycles[4], the ambiguity in defining system autonomy, machine automation and human interaction contributes to alternate architecture assessment and trade studies leading to ambiguous requirements development.

A. *Defense Science Board Task Force on Autonomy*

The Defense Science Board (DSB) Task Force on Autonomy[4] 'reviewed many of the DoD-funded studies on "levels of autonomy" and concluded that they are not particularly helpful to the autonomy design process. These studies attempt to aid the development process by defining taxonomies and grouping functions needed for generalized scenarios. They are counter-productive because they focus too much attention on the computer rather than on the collaboration between the computer and its operator/supervisor to achieve the desired capabilities and effects. Further, these taxonomies imply that there are discrete levels of intelligence for autonomous systems and that classes of vehicle systems can be designed to operate at a specific level for the entire mission.'

The DSB was asked to study relevant technologies, ongoing research and the current autonomy-relevant plans of the Military Services, to assist the USDoD in identifying new opportunities to more aggressively use autonomy in military missions, to anticipate vulnerabilities and to make recommendations for overcoming operational difficulties and systemic barriers to realizing the full potential of autonomous systems.

The DSB has concluded that autonomy technology is being underutilized as a result of obstacles within the USDoD inhibiting the acceptance of autonomy and unmanned systems. Key among these obstacles are a) poor design, b) lack of effective coordination of research and development and c) insufficient resources or time to refine concepts of operations[4].

The DSB states that 'Autonomy is a capability (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, "self-governing." Unfortunately, the word "autonomy" often conjures images in the press and the minds of some military leaders of computers making independent decisions and taking uncontrolled action. While the reality of what autonomy is and can do is quite different from those conjured images, these concerns are in some cases limiting its adoption. It should be made clear that all autonomous systems are supervised by human operators at some level and autonomous systems' software embodies the designed limits on the actions and decisions delegated to the computer…Instead of viewing autonomy as an intrinsic property of an unmanned vehicle in isolation, the design and operation of autonomous systems needs to be considered in terms of human-system collaboration.'

To address the issues that are limiting more extensive use of autonomy in USDoD systems, the DSB recommends [4] a crosscutting approach that includes the following key elements:

- The DoD should embrace a three-facet (cognitive echelon, mission timelines and human-machine system trade spaces) autonomous systems framework to assist program managers in shaping technology programs, as well as to assist acquisition officers and developers in making key decisions related to the design and evaluation of future systems.
- The Joint Staff and the Military Services should improve the requirements process to develop a mission capability pull for autonomous systems to identify missed opportunities and desirable future system capabilities.

B. *Mathematical Representation of System Autonomy*

A system autonomy assessment tool must show a mathematical relationship between human interaction and machine automation [4]. Being a software only model, this tool would be a good candidate for the AGILE development methodology. A workable and measurable definition of system autonomy (SA) is then defined as a functional of human interaction (HI) and machine automation (MA):

$$SA = F[MA, HI] \quad (1)$$

If System Autonomy is considered as a vector, then the relationship between HI and MA would provide the scalar component. Mathematical assessment of SA as a vector representation is far more logical than using discrete integer levels.

The many unmanned air vehicles requires different levels of human interaction and supervisory control. Unmanned Air Vehicles range in sophistication and may need one or more human supervisors to successful carry out a surveillance mission. Equation 1 describes a single operator, single UMS configuration; the SA function from equation 1 above is modified as follows:

$HI = G[HI_1, HI_2,\ldots HI_n]$, where n is the operators needed during the mission (2)

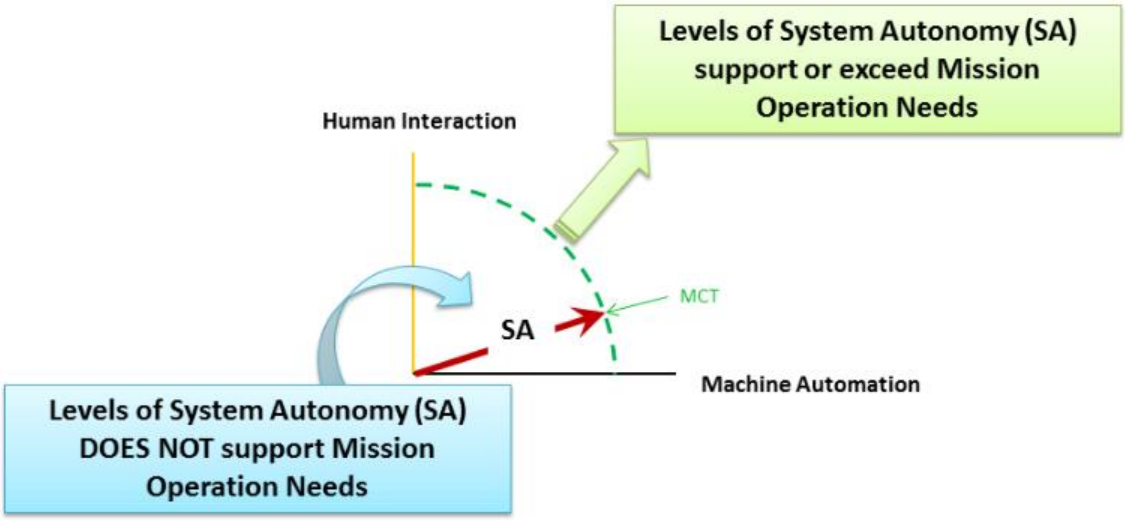


Fig. 2. System Autonomy Trade Space

An alternative design for the system autonomy equation is a network of unmanned vehicles controlled by a single operator. This increases machine automation and facilitates a network of multiple UMS, operating concurrently and is supervised by a single controller. The alternate design would have multiple UMS operating sequentially and supervised by a single controller.

As the number of intelligence, surveillance and reconnaissance (ISR) missions increase, a single operator would control multiple UMS and in this scenario the equation is modified as:

$MA = K[MA_1, MA_2 \ldots MA_m]$, where m is the number of UMS (3)

1) System Autonomy as a Vector

Fig. 2 graphically depicts System Autonomy as a vector in the MA/HI trade space. The magnitude of the SA vector is determined from the contributions of MA and HI component variables. The magnitude indicates whether the system architecture would meet mission objectives. The significance of the angle is discussed later. When the required value of the vector SA is set to a constant throughout the trade space, this defines the minimum autonomy levels needed to meet mission requirements. The dotted arc represents the Minimum Capability Threshold (MCT) where SA would meet this threshold. If the magnitude of the candidate system vector fell short of the MCT, then some mission objectives would not be accomplished.

Additional contributions from MA and HI would be needed to increase the magnitude of the SA vector. Magnitude exceeding the MCT indicates more than needed system autonomy to execute the mission. Normalizing the SA vector to a value of one (SA=1) allows further investigation to the relationship between MA and HI. Setting the SA vector to intersect with the HI axis sets the value for MA = 0 and HI = 1. This represents complete machine dependence on human interaction. Setting the SA vector to intersect with the MA axis sets the value for MA = 1 and HI = 0. This represents complete machine independence from human interaction. Maintaining SA=1 as the vector moves within this plane scribes the MCT arc and provide the mathematical relationship between SA, HI and MA. This spare capacity can be viewed as capability reserves or targeted for reduction as potential life cycle cost efficiencies. The magnitude of the vector becomes

$$SA = \sqrt{(HI^{\wedge 2}+MA^{\wedge 2})} = F[HI,MA] \quad (4)$$

This allows the relationship between MA and HI to be defines as:

$$HI=\sqrt{(1+MA^{\wedge 2})} \quad (5)$$

Treating SA as a vector allows for analysis of candidate systems during the AoA and concept of operation activities where the systems design is developed.

2) SA Phase Angle

The angle Ψ, Fig. 3 provides an indication of the technology inherent in the configuration. The angle, with the scalar magnitude, describes SA as a vector. This allows vector mathematics when assessing system of autonomous system configurations. The angle is expressed as:

$$\Psi= \tan^{-1}[MA/HI] \quad (6)$$

The SA phase angle provides a relative comparison of the technology base for the candidate system. The smaller the difference in angles indicate that the candidate systems share the similar technology architectures and comparative analysis is relative straight forward. The greater the difference between the phase angles indicates that the systems have a diverse technological base making any comparison more complex.

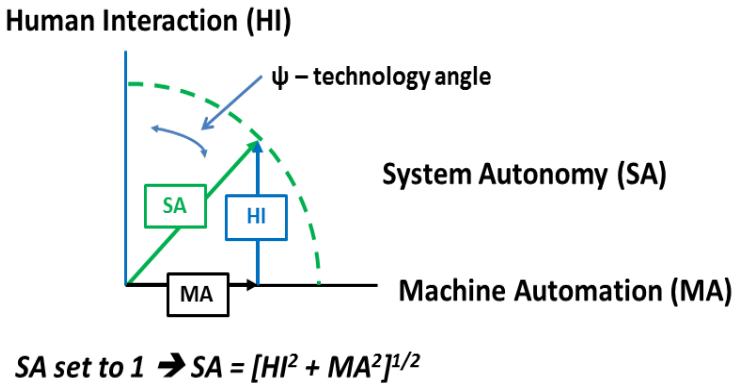


Fig. 3. Technology Angle

$$\overline{|SA|} \stackrel{def}{\cong} \sqrt{(MA^2 + HI^2)}\ and\ \tan^{-1}(MA/HI) \qquad (7)$$

3) System Autonomy Trade Space

Expressing identifies the magnitude and phase angle of the vector. This provides the algorithmic assessment capabilities the current methods cannot provide.

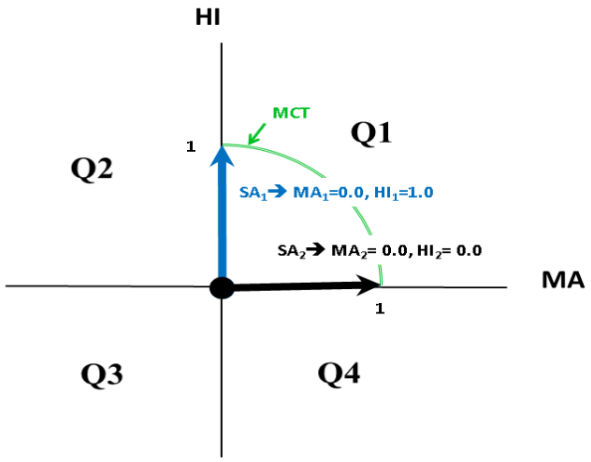


Fig. 4. Diverse Technologies

Fig. 4 shows two systems of equal magnitude. Both systems meet the MCT but the difference between the Ψs is 90° and indicates an extreme divergence of technologies. One system is tele-operated, Level 1. The other system exhibits android behavior, Level 10 and does not depend on any human interaction [28]. A comparison between the two systems architectures would not be straightforward because they operate in significantly diverse manners. For clarity, a systems candidate is shown in quadrant (Q1) unless uncooperative assessments are needed. Systems in different quadrants have vector components that would tend to negate, resulting in a smaller magnitude value. Systems in Q3 would be considered as countermeasures to systems in Q1 and are diametrically opposing forces. Systems in Q2 and Q4 have utility and assessments that may include fault, stress test or destabilizing scenarios. Future missions would include collaborative operations of more than one unmanned vehicle and the equation would be expanded to have two or more unmanned systems, UMS;

$$SA = F[SA_1] + F[SA_2] + \ldots F[SA_k] \quad \text{where } k \text{ is the number of UMS} \qquad (8)$$

Collaborative missions would include mixed UxV modes such as surface (USV), ground (UGV), air (UAV) and underwater (UUV) contributions. In the above relationship, UxV would be substituted by the appropriate type and number of UAVs, USV, UGVs and or UUVs as identified by the mission requirements. If one operator controls multiple UxVs, then the variable permutations of this model grow in complexity and a clear need for a model and methodology during AoA and CONOPS development becomes evident. The multiple combination UMS equation for SA becomes:

$$SA = F[SA_{UAV}] + F[SA_{UGV}] + F[SA_{USV}] + F[SA_{UUV}] \qquad (9)$$

The additive effect of SA from multiple subsystems is further described and depicted in section 0. Inclusion of dynamic variables like mission difficulty, meteorological impacts and many other probabilistic variations just increases the complexity of understanding and defining requirements.

4) Contextual System Autonomy

In previous sections, System Autonomy was discussed as a two dimensional vector. In more representative scenarios, system autonomy, human interaction and machine automation vary throughout the mission. Varying machine automation to meet mission needs is currently possible by commanding the machine to perform less than its maximum design capabilities allow. In some cases, new software can be downloaded to perform more efficiently. If the capability is not mechanically inherent in the machine, hardware reconfiguration by the machine itself is not supported by current technologies. The same may not be true of the human interaction element.

Fig. 5 provides time as the third dimension to the trade space. Expanding the trade space to a third dimension should not infer a three dimensional SA vector. Instead the magnitude of the two dimensional SA vector is plotted against the third axis which represents the mission time. The mission phases may evolve and require a change from one type of UMS to another or a change of operator skills. In this case a Mission Phase would describe the system autonomy needed to conduct the mission phase peculiar activities. Mission Phase changes can appear as discontinuities in the SA level.

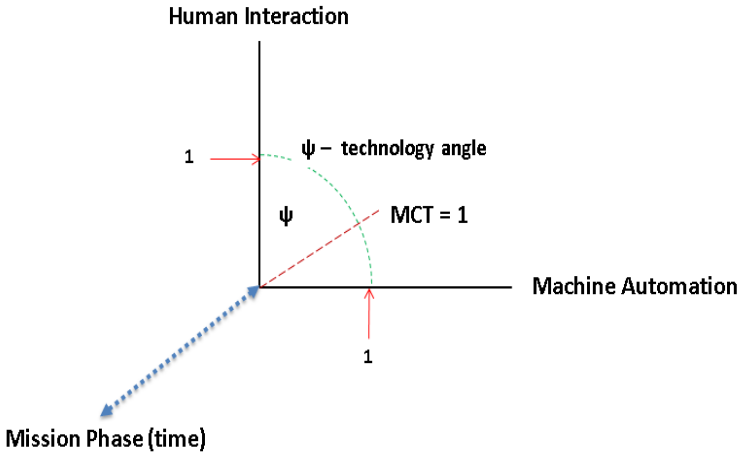


Fig. 5. Contextual Autonomy

5) A Dependency on Mission Phase

Time becomes a consideration in two ways. Complex mission scenarios may require several changes of system autonomy levels due to the changing phases of the mission like transit and area surveillance. This causes the SA vector to have a Mission Phase dependency. The combination of human supervision/interaction and the level of needed machine automation may need to vary within each discrete mission phase. This causes the SA Vector to have a time variant dependency.

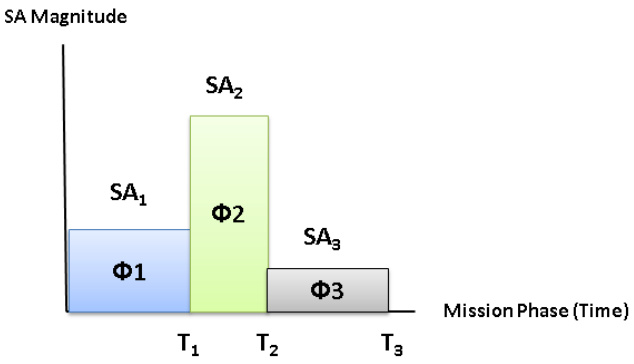


Fig. 6. Mission Phase

Fig. 6 graphically represents the two dimensional SA vector throughout the notional mission duration. In this example, mission scenario with three phases – Φ1 is the transit to operation area, Φ2 is the surveillance and reconnaissance and data gathering activity and Φ3 is the return transit. Each of these mission segments may require a specific level of SA. In this depiction, each SA is constant through the mission phases. This is not typical and most often observed is that there is some SA level variability with each mission phase.[3]

SA Dependency on Time

[3] This is the author's observation in working with ISR UUVs, Anti-Torpedo-Torpedo, Littoral Combat Ship Mission Packages, several Mine Neutralization UUS and missile and torpedo programs.

System Autonomy or the Human Interaction can be a dynamic within a Mission Phase, Fig. 7. As an example, real time video could be collected during surveillance activity. The Human Interaction could be higher at the start of the data collection run as processor settings may need to be changed to accommodate the environmental conditions. The Human Interaction could be reduced during the data collection run and increase again at the end to verify data collection and processing. Human Interaction could also be a function of false positives that need to be interpreted and discounted.

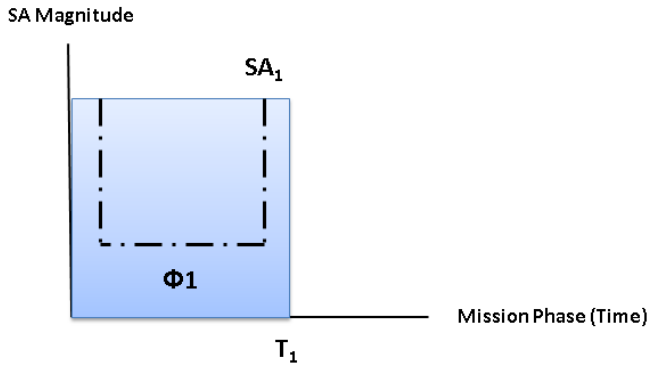


Fig. 7. Time Dependency

SA can be a linear or piece-wise aggregate of discrete action. As an example, during the launch phase, human interaction could be high at the very beginning, lower during system built in testing and high again just as the vehicle launches and separates from the cradle/gantry.

The SA and HI can be variable, distinctive and different for each Mission Phase. In the event that the system is composed of subsystems, then combined SA or HI is the accumulation of the individual subsystem contributions. The cumulative SA levels contributions from the subsystem involved in each Mission Phase is graphically shown in, Fig. 7.

III. Comparative Autonomy Assessments

In 2007, Southwest Research Institute applied the Autonomy Levels For Unmanned Systems (ALFUS) framework to assess the achieved autonomy levels of eight

unmanned ground vehicles (UGV) [23] [4]. ALFUS is a framework that has been developed by a consortium of government and non-government agencies during several workshops [18][19]. Note that in the ALFUS methodology, the variable HI is Human Independence not Interaction as used in this paper. The UGVs were categorized into four groups by market area and use. This allowed some narrowing of the definitions of Mission Complexity (MC) and Environmental Difficulty (ED) within each group. Even with this pre-filtering, some ambiguities in assessment existed and straight comparisons outside of the grouping are not straightforward. Within the Passenger Vehicle grouping, both UGVs require a human operator. The Human Operator in NavLab actuated the throttle and brake but not steering. The Operator in the NavLab had to monitor the UGVs unsafe lane positions and distances to vehicles. The ALFUS HI does not portray the involvement of the human operator as he would be required 100% of the time during these tests. Using the ALFUS methodology, SRI assessed the MC, ED, HI and ∑ for each UGV within the specific group constraints. In the ALFUS methodology, MC and ED seem to have some overlapping definitions. The mission complexity includes terrain and hostilities in the case of the Military grouping which spills into the environmental difficulty which also take into account terrain and hostilities. When SRI summed the three variables, the ALFUS autonomy assessments of the UGVs were very similar.

In the ALFUS methodology, the variable HI provides for human independence and not Human Interaction. If this variable is viewed as a form of machine automation (MA), then the algorithmic assessment can be applied, **Error! Reference source not found.**. The algorithmic value of human interaction (HI and the technology base angle (Ψ) are calculated (normalized) for each UGV using equations 5 and 6. Although human operators were needed to operate some of the UGVs, no adjustments to the ALFUS derived levels were included. In the case of the NAVLAB UGV, the researched operated actuated the throttle and brake manually thereby increasing the human interaction to a greater level then indicated. When the UGVs are further segregated into subgroups, the algorithmic assessment in**Error! Reference source not found.** Algorithmic (1) shows that the technology bases of the UGVs within each grouping maybe too diverse for straight comparisons. This is evident within the Passenger Vehicle category. Manual categorization into categories is not sufficient for assessment of system autonomy of between candidates of an AoA.

Comparisons of those UGVs with similar Ψs are straight forward and other factors such as life cycle cost can compared. Performance attenuating parameters such as terrain difficulty or hostilities can be applied in stochastic studies in developing concept of operations. The ALFUS methodology provides a combined label assessment and parametric sensitivity studies could not be performed easily.

In the analysis performed by SRI, the three ALFUS variables were summed and identified as Algorithmic (2) in the table. An alternate assessment of MA is done if the ALFUS variables are averaged and then applied as MA in a similar fashion done by SRI. As was found in the SRI assessments, the Ψs of the UGVs become numerically closer, indicating relative straight forward comparisons are possible. This could potentially increase the number of candidate systems during system requirement and AoA developments, not possible with current assessment methodologies and tools. As in the previous case, factors such as life cycle cost can be included for comparison. Parametric sensitivities and stochastic modeling can be performed to contribute to AoA, CONOPS and requirement development not possible with the ALFUS framework. Summarizing, a label based system autonomy tool has very limited usefulness in defining and developing system concepts. Label assessment tools do not provide visibility into system components or design contributors. Label assessment tools do not support parametric sensitivity or stochastic analyses. An algorithmic assessment tool can support design activities in developing system concepts. This is the inference reached by the DSB[4].

IV. Summary and Conclusions

Autonomous systems result from a complex integration of human intelligence supervising machine automation to adapt to unforeseen events encountered during operations Although significant work has been undertaken, conventional SA assessment frameworks are not suited for trade studies in support of AoA, CONOPS and requirements development. Missions are becoming more complex and require ever-increasing capabilities to adapt to varying unknown situations. Autonomy is a complex function of many dynamic and widely varying parameters and requires a mathematical relationship between Human Interaction and Machine Automation to provide the design tradeoff study capabilities needed during early development phases. The Defense Science Board stated that machine automation and human interaction assessments need algorithmic solutions instead of the label methodology.

TABLE I. Algorithmic Assessment

		ALFUS				Algorithmic (1)			Algorithmic(2)		
Category	UGV	MC	ED	HI*	Σ	MA	HI	Ψ	MA	HI	Ψ
Passenger Vehicles	NavLab	4	7	6	17	0.6	0.8	36.870	0.567	0.824	34.518
	ARGO	4	7	8	19	0.8	0.6	53.130	0.633	0.774	39.296
Transit & Frieght	CMU Houston Metro Bus	5	3	8	16	0.8	0.6	53.130	0.533	0.846	32.231
	CyberCars	6	5	10	21	1	0	90.000	0.700	0.714	44.427
ET Rover	Spirit	6	7	6	19	0.6	0.8	36.870	0.633	0.774	39.296
Military	XUV DEMO III	6	6	9	21	0.9	0.436	64.158	0.700	0.714	44.427
	Crusher	7	7	7	21	0.7	0.714	44.427	0.700	0.714	44.427
DARPA Grand Challenge	Stanley	4	6	10	20	1	0	90	0.667	0.745	41.810

[4] In the SRI paper, conflicting values for HI were given at 8 and 10 for the Houston Metro Automated Bus. The value of 8 was used for HI as this seemed to be consistent with later calculation made in the paper.

The mathematical relationship described in this paper provides a basis for such a framework. Incremental and partial capabilities models can be developed using rapid design methodologies. Future development of System Autonomy Assessment tools would provide additional capabilities and mature the requirements refinement process for the development of autonomous systems currently not available.

ACKNOWLEDGEMENTS

Mr. William Glenney, US Naval War College, for providing valuable access the CNO Strategic Studies Group report on the future of US Navy Unmanned Vehicles. Mr. Steve Koepenick, past Director of Association for Unmanned Vehicle Systems International (AUVSI) and past Executing Agent for Department of Defense Unmanned Autonomous Systems initiatives, for providing access to the many working groups; materials and reports on System Autonomy.

REFERENCES

[1] Acquisition ASotAFf. Early Systems Engineering Guidebook. In: Force USA, editor.: Assistant Secretary of the Air Force for Acquisition; 2009.

[2] Ambler S, Holitza M. Agile for Dummies, IBM Limited Edition. Hoboken, NJ: John Wiley & Sons, Inc.; 2012.

[3] Bergey JK, Blanchette Jr. S, Clements PC, Gagliardi MJ, Klein J, Wojcik R, et al. U.S. Army Workshop on Exploring Enterprise, System of Systems, System, and Software Architectures. Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute; 2009.

[4] Board DS. TASK FORCE REPORT: The Role of Autonomy in DoD Systems. Washington DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics; 2012.

[5] Boehm B. Spiral Development: Experience, Principles, and Refinements. In: Hansen WJ, editor. Spiral Development Workshop February 9, 2000. Pittsburgh, PA 15213-3890: Carnegie Mellon SEI; 2000.

[6] Clough BT. Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway? In: Laboratory AFR, editor. Proceedings of the Performance Metrics for Intelligent Systems Workshop. Gaithersburg, Maryland: Wright-Patterson AFB; 2002. p. 7.

[7] Curtin NP, Francis PL. Major Management Issues Facing DOD's Development and Fielding Efforts. In: Office USGA, editor.2004.

[8] Defense So. 2014 Quadrennial Defense Review. In: Defense Do, editor. Washington DC: Departmenmt of Defense; 2014.

[9] DEYST JJ, EGAN JF. AUTONOMOUS VEHICLES IN SUPPORT OF NAVAL OPERATIONS: THE NATIONAL ACADEMIES PRESS, Washington, D.C.; 2005.

[10] Director SaSE. Systems Engineering Guide for Systems of Systems. In: Office of the Deputy Under Secretary of Defense for Acquisition and Technology SaSE, editor. Washington D.C.: Office of the Deputy Under Secretary of Defense for Acquisition and Technology, Systems and Software Engineering; 2008.

[11] Douglass PhD BP. Agile Systems Engineering. Innovate2012: IBM Software; 2012.

[12] Elmendorf D. Long Term Implication of the 2014 Future Years Defense Program. In: Office CB, editor. Washington DC: Congressional Budget Office; 2013.

[13] GORTNEY WEV, USN. JOINT CAPABILITIES INTEGRATION AND DEVELOPMENT SYSTEM. In: Defense Do, editor. Washington DC10 January 2012.

[14] Hansen EC. A Relationship Approach to Autonomy Metrics. AUVSI North America 2011. Washington, DC2011. p. 14.

[15] Haven KP. 'Danger, Will Robinson!' - Controlling the Public Perception of Unmanned Systems and Robotics. Unmanned Systems. 2011:4.

[16] Hoffman M. US Army acquisition frustration spills into open forum. C4ISR Journal. 2011;10:1.

[17] Hogg JRAr. CNO Strategic Studies Group XXVIII, The Unmanned Imperative. In: Navy US, editor. December 2009 ed. Newport: Navy War College; 2009.

[18] Huang H-M. Terminology for Specifying the Autonomy Levels for Unmanned Systems: Version 1.0. In: (U.S.) NIoSaT, editor.: US. Department of Commerce; 2004.

[19] Huang H-M, Messina E, Albus J. Toward a Generic Model for Autonomy Levels for Unmanned Systems (ALFUS). In: Division NIoSaTIS, editor. Performance Metrics for Intelligent Systems (PerMIS) Workshop. Gaithersburg, MD2003.

[20] Iannota B. Staying Focused on Automation. C4ISR Journal. 2011;10:1.

[21] IBM. Agile in the Embedded World. UBM Tech, a division of United Business Media LLC. ALl Rights Reserved.; 2013.

[22] Jean GV. Army deploying robotic MULE to troops in Afganistan. National Defense. 2011;XCVI:1.

[23] McWilliams GT, Brown MA, Lamm RD, Guerra CJ, Avery PA, Kozak, Kristopher C. , et al. Evaluation of Autonomy in Recent Ground Vehicles Using the Autonomy Levels for Unmanned Systems (ALFUS) Framework. Washington DC: Southwest Research Institute; 2007.

[24] Murphy PhD R, Shields J. The Role of Autonomy in DoD Systems. In: Defense Do, editor. WASHINGTON, DC 20301â€"3140: OFFICE OF THE SECRETARY OF DEFENSE; 2012.

[25] National Research Council (U.S.). Committee on Autonomous Vehicles in Support of Naval Operations. Autonomous vehicles in support of naval operations. Washington, D.C.: National Academies Press; 2005.

[26] North Atlantic Treaty Organization. Research and Technology Organization. Systems Concepts and Integration Panel. Integration of systems with varying levels of autonomy. Rto Tr-Sci-144. Neuilly-sur-Seine Cedex, France: North Atlantic Treaty Organization, Research and Technology Organization; 2008.

[27] O'Rourke R. Unmanned Vehicles for U.S. Naval Forces: Background and Issues for Congress. In: Congressional Research Service TLoC, editor.: The Library of Congress; 2006.

[28] Parasuraman R, Sheridan TB, Wickens CD. A Model for Types and Levels of Human Interaction with Automation. IEEE Trans Syst Man Cybern B Cybern. 2000;VOL. 30:12.

[29] Pernin CG, Axelband E, Drezner JA, Dille BB, Gordon IV J, Held BJ, et al. Lessons from the Army's Future Combat Systems Program. In: Corporation. R, editor.March 4, 2013.

[30] Pontius RW. Acquisition of Information Technology Improving Efficiency and Effectiveness in Information Technology Acquisition in the Department of Defense. In: Defense Do, editor.2012.

[31] Royce WW. Managing the Development of Large Software Systems. IEEE WESCON1970.

[32] Stone M. Brief on Autonomy Initiatives in the US DoD. In: Defense Do, editor.2012.

[33] Technology OotDUSoDfAa. Systems Engineering Guide for Systems of Systems. In: Technology OotDUSoDfAa, editor. Washington, DC: ODUSD(A&T)SSE; 2008.

[34] Tiron R. Army to end robotic vehicle, aircraft efforts. The Hill2010.

[35] Under Secretary of Defense for Acquisition TaL. DoD Instruction 5000.02. In: Defense Do, editor. Washington DC2013.

[36] University DA. Defense Acquisition Guidebook. In: University DA, editor.: Defense Acquisition University; 2012.

[37] Valavanis KP. Advances in Unmanned Aerial Vehicles State of the Art and the Road to Autonomy. Intelligent Systems, Control and Automation: Science and Engineering 33. Dordrecht: Springer, SpringerLink (Online service),; 2007.

[38] van der Vyver JJ, Christen M, Stoop N, Ott T, H. SW, R. S. Towards genuine machine autonomy. Elsevier Science. 2003;8 December 2003.

[39] Westermann WE. A METHODOLOGY AND MODEL TO DEVELOP COMPLEX SYSTEMS FROM REQUIREMENT NETWORKS [Dissertation]. Dallas, TX: Soutnern Methodist University; 2008.

[40] Whittaker W. High performance robotic traverse of desert terrain. Washington, D.C., Oak Ridge, Tenn.: Sandia National Laboratories, United States. Dept. of Energy, United States. Dept. of Energy. Office of Scientific and Technical Information, United States. Dept. of Energy ; distributed by the Office of Scientific and Technical Information, U.S. Dept. of Energy; 2004.

Prototype of a Web ETL Tool

Matija Novak, Kornelije Rabuzin
Faculty of Organization and Informatics
University of Zagreb
Varazdin, Croatia

***Abstract*— Extract, transform and load (ETL) is a process that makes it possible to extract data from operational data sources, to transform data in the way needed for data warehousing purposes and to load data into a data warehouse (DW). ETL process is the most important part when building the data warehouse. Because the ETL process is a very complex and time consuming, this paper presents a prototype of a web ETL tool that offers step-by-step guidance through the entire process to the end user. This ETL tool is designed as a web application so users can save time (and space) required for installation purposes.**

Keywords—ETL; data warehouse; web; ETL tool

I. INTRODUCTION

Databases (DB) have been used for many years and it is hard to imagine any (transaction) application that wouldn't use some database. Over time people realized that databases, although they support daily operations, are not good source when complex analysis must be made on data. Merging data from multiple tables, the complexity of the model (as such), the inability to generate reports by end users and (in)effectiveness of such approach resulted with the need to reorganize (transform) data into a form that will be suitable for analysis. This form is called a data warehouse [1, p. 85].

The basic idea of data warehouses is to store data in such a way that users can understand and analyze data. R. Kimball and J. Caserta define the data warehouse as follows: "A data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making." [2, p. 23]. This definition says that data warehouses are used to support decision-making. The data warehouse would not be good without the iterative process of extracting, cleaning, conforming and loading data (the so called ETL process) from various sources into the star schema model.

When we talk about data organization in the data warehouse, we distinguish between fact and dimension tables. While dimension tables contain large number of attributes that we use when analyzing (filtering) data, fact tables contain measures to quantify business processes (number of product units sold, number of orders, number and duration of calls, etc.). For end users such model is understandable and they can independently create necessary reports.

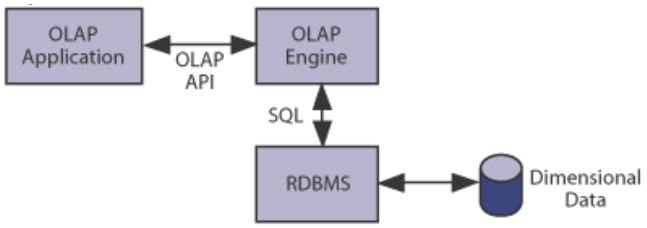


Fig. 1. ROLAP model [3]

Basically there are two mechanisms (ways) that can be used to store data in the data warehouse (Fig. 1): relational online analytical processing (ROLAP) and multidimensional online analytical processing (MOLAP) [4, p. 165] While ROLAP stores data in tables, MOLAP stores data in special structures also known as cubes. There are advantages and disadvantages but we will not discuss them in this paper (more can be found in [9]).

If one looks at today's market, one can find various Business Intelligence (BI) tools that are used to produce reports by using data from data warehouses. Although data warehouses are very valuable sources of data, the main problem in the construction of the data warehouse is the so called ETL process. Systems for extraction, transformation and loading of data (ETL systems for short) are the foundation of data warehouses. When constructing a data warehouse 70 percent of time and resources is used for the ETL purposes (by Inmon 80 percent [5, p. 295]).

Building a data warehouse is expensive, time consuming and complex job and the ETL phase is the most critical one. Because of that the idea of this paper is to present the ETL tool that should facilitate and accelerate the process of ETL. This ETL tool offers the user step-by-step guidance through the entire ETL process. In addition this ETL tool is designed as a web application and users can save time (and space) required for installation. This tool can start from heterogeneous sources of data and result is a dimensional model stored in a relational database which can be used for other purposes (primarily for building reports by means of some BI tool).

This paper is structured as follows: the second section describes the related work and the third section the basics of ETL. Next, the model of the ETL tool is shown and several screen shoots are given. In the end of the paper some open questions are addressed (future work) and the conclusion is presented.

II. ETL TOOLS

There are various professional tools, which can be used to assist user in the ETL process; however, the problem of these tools is their complexity and/or price. For example, if we take free tool Talend Open Studio, its features are great and user can execute complex operations. But, the tool can be very difficult and confusing, especially if the user is not familiar with the ETL process. Because the tool has a number of possibilities, it is necessary to examine what our individual elements (or rather objects) allow us to do and what are their attributes.

In addition, there is service-oriented architecture (SOA) ETL Framework described in [7] that tries to split the tightly coupled functionalities of an ETL tool into separate parts that can be used as services.

To the authors knowledge there is no such thing as a completely web based tool that would integrate the learning of the ETL process into the tool itself. Furthermore, the ETL tool described in this article is completely web based (it can be easily accessed through web browser, no installation is needed and multiple users can use it at the same time). To avoid problems with the ETL process, the created tool guides the user through the ETL process and teaches him during the way; so the basic idea is that it can be used by people not familiar with the ETL process.

III. ETL

The ETL process is a set of activities that are not visible to the end user and that are taking place in the background. In addition to retrieving information from different sources, many activities need to be performed on data [2, p. xxi]: mistakes have to be corrected, data needs to be structured, etc.

The ETL process (Fig. 2) has three steps [6, p. 139]:

- *Data extraction* – accessing data sources in order to retrieve (required) data.
- *Data transformation* - In this step data collected from various sources is checked, cleaned and conformed, i.e. data undergoes a series of activities in order to improve the quality of data. [4, p. 375]
- *Data loading* - extracted and transformed data is loaded into the data warehouse (dimension and fact tables).

While extraction and loading only transfer data, transformations are really changing data. Kimball and Caserta propose the so-called Extracting, Cleaning, Conforming, and Delivering (ECCD) instead of the ETL, but either way in the end data has to be loaded in the data warehouse. ECCD consists of four steps [2, pp. 18-19]:

- *Extraction* – the first step is to take data from different sources and store it in the ETL environment in order to make the necessary processing.
- *Cleaning* – performing the first transformation of data in order to enhance the quality of the original data.
- *Conforming* – This step is necessary if there are two or more data sources. Various sources tend to have

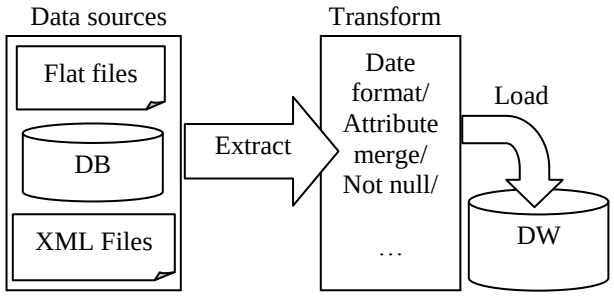


Fig. 2. ETL process steps

differently shaped and stored data and there is a need to synchronize data (resolve conflicts when different names are used, resolve the problem of duplicates, etc.).

- *Delivery* – The last step is the same as for the ETL (loading data into the data warehouse). This process of loading data can be further divided into two parts [2, pp. 161-254]:
  - Loading data into *dimension tables* – they contain information that allows understanding (interpreting) the data in fact tables
  - Loading data into *fact tables* – central tables that contain numerical values.

There is also the fifth step (the so called management) which is not part of the flow of data processing, but it is used for system and process management of the ETL environment. ETL and ECCD describe the same data processing activities and the end result is the same, yet the ECCD is somewhat better because the steps define in detail the activities to be carried out in the processing of the original data and it separates activities related to a single source of data and activities that include multiple data sources. Nevertheless the term ETL is so "domestic" that it is not reasonable to expect that it is replaced in the near future.

A. Metadata

During the ETL process various metadata is generated. The ETL metadata is divided into four main categories [2, pp. 367-368]:

- *ETL job metadata* – is a container of transformations that manipulate the data. Every ETL task is captured here.
- *Transformation metadata* – contains information about every transformation that is used inside of the ETL jobs.
- *Batch metadata* – in the ETL process batches are used to run collections of jobs together. Batches can contain sub-batches and schedules can be made to run batches periodically. All that information is stored in batch metadata.
- *Process metadata* – is generated when batches are executed. Process metadata has information on whether loading of data (into the DW) was successful or not.

1) Logical data map

At the beginning of the ETL process, it is necessary to make a logical data map. The logical data map documents the links between the columns (fields) in the source and the columns in the destination table (in the data warehouse). Logical data map is one of the most important and most useful metadata generated by the ETL. Header of the logical data map is shown in the following table [2, pp. 56-71].

Once created, the logical data map provides information about what needs to be extracted, from where, how to process data and where it needs to be saved after processing. The logical data map is useful throughout the entire ETL process.

TABLE I. HEADER OF THE LOGICAL DATA MAP [2, P. 60]

Target					Transformation
Table name	Column name	Data type	Table type	SCD type	
Source					
Database name	Table name	Column name	Data type		

2) Data sources

A data warehouse often uses different data sources (Enterprise Resource Planning (ERP) Systems, extensible markup language (XML) files, databases and flat files). No matter which source is used, specific metadata is required. The following metadata attributes are minimally required [2, p. 362]:

- *Database or file system* – "The name commonly used when referring to a source system or file." [2, p. 362]
- *Table specification* – "The ETL team needs to know the purpose of the table, its volume, its primary key and alternate key, and a list of its columns." [2, p. 362]
- *Exception-handling rules* – Necessary information related to the quality of data and how should the ETL process manage them.
- *Business definitions* – It's good to get the business definitions as these two or three sentences are very useful when you need to understand data.
- *Business rules* – "Every table should come with a set of business rules. Business rules are required to understand data and to test for anomalies." [2, p. 362]

Types of data sources can be:

- *Flat Files* - In most data warehouses regular files can't be avoided. Flat files can be used in the ETL system for at least three reasons [2, pp. 90-91]: *delivery of source data*, *working/staging tables* or *preparation for bulk load*. There are two types of files [2, pp. 91-93]: *fixed length flat files* and *delimited flat files*.
- *XML files* - In recent years the XML is used very much. XML files are good for the ETL process because they are self-documented unlike ordinary files that are not. XML files are often used for data exchange and provide independence from the specific computational implementations [4, p. 126].
- *Operational databases* - the most common source of data for the data warehouse. Benefits of databases regarding the ETL phase are [2, pp. 40-41]: *Apparent metadata*, *Relational abilities* (exp. referential integrity), *Open repository* (data can easily be accessed by any structured query language (SQL) compliant tool), *DBA Support* (there is a group responsible for data in database management system (DBMS)), *SQL interface*, *etc*.
- Other sources:
  - *ERP Systems* – systems that are quite common in organizations.
  - *Master data management (MDM) Systems* - are centralized resources designed to hold the main copy of the key entity, such as a customer or a product.
  - *Web log* – for example a control document that is automatically created from the Web server.

IV. THE MODEL OF THE ETL TOOL

The following figure (Fig. 3) shows the high level architecture of the proposed ETL tool. The user uses web interface to define the metadata (i.e. user creates project, process, group, destination, etc.) that the ETL processing will use. When all data is entered, user runs the thread that extracts information from one source, then performs defined transformation (as necessary) and finally loads data into the data warehouse. After one source is completed, the thread proceeds to the next source. Possible improvement is to implement multithreading in order to process multiple sources at once.

A. ETL thread

Data processing is made by the thread that starts after the metadata is entered. Fig. 4 shows the class diagram of this part of the tool. When you start the thread class "Main logic", it is instantiated and it then instantiates classes "Extraction", "Transformation" and "Load". After that the methods of the class "Load" are called to create the destination (dimension and fact tables). Then, the logical data map is read and information is stored into two vectors. The first vector contains metadata relating to data for dimension tables and second vector stores data for fact tables. The thread then moves and processes dimensions, one by one, and SQL query for extraction is created and run. After that, data is transformed as it is described in the metadata entered by the user; after the transformations are done, the loading starts to load data into the data warehouse (row by row). When dimensions are finished, the fact tables are processed in the next step (the procedure is the same but one has to have in mind that fact tables have to be connected to specific dimension tables).

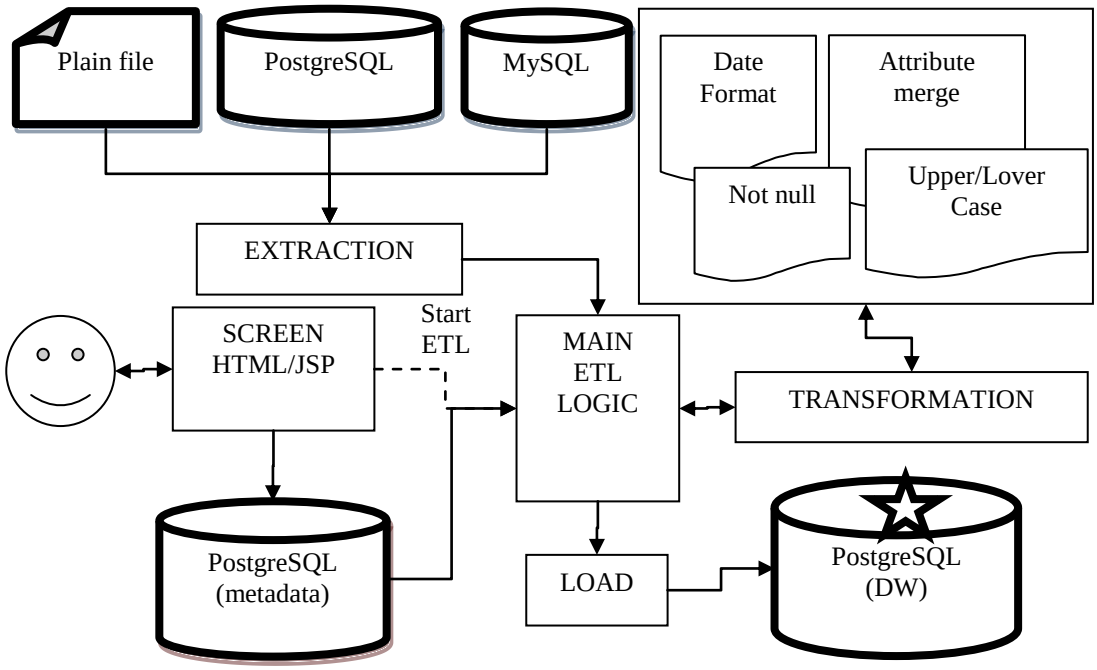


Fig. 3. ETL tool High-level architecture

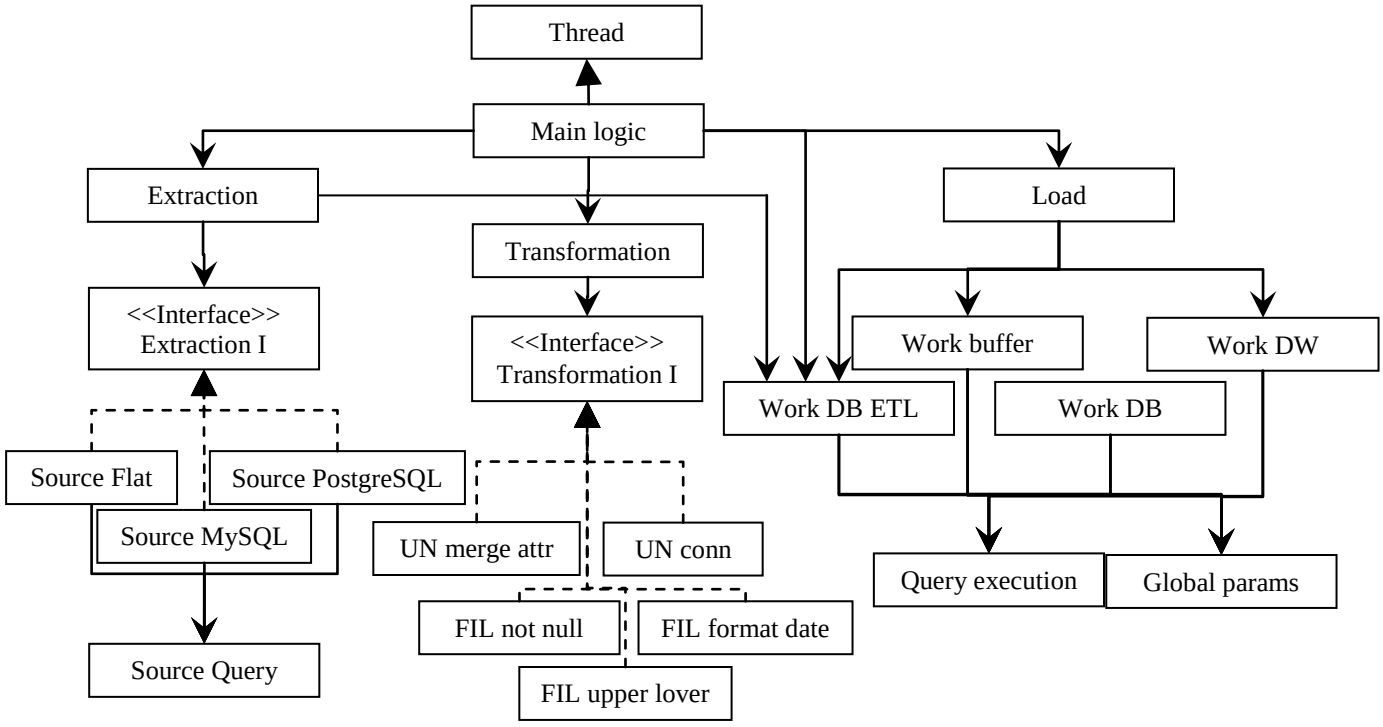


Fig. 4. Class diagram – ETL thread

B. Dynamic loading

In order to create a flexible tool and have the option of upgrading, dynamic load of classes and JSP files has been implemented in two places:

- *In source extraction part* – for every source type one class has been made;
- *In data transformation part* - for every transformation that the tool can perform one class has been made;

Since each type of source and each transformation have their own class, it is possible to add new types of sources or new transformations. All you need to do is create a class (and if needed a JSP file) and add metadata info about it.

Three important things enable dynamic loading of classes:

- Each class of source type (or transformation) must have a method that returns an instance of a class within the class itself (Fig.5).
- The interfaces implemented by source type or transformation classes (Fig. 6).
- The class that has methods to search for required class through its name and dynamic load of the class into memory and methods to search for functions within the retrieved class that return an instance of the desired class (Fig. 7)

In addition to the dynamic class loading, transformations also use dynamic load of JSP files which contain fields (if necessary) that the user must fill in when choosing this particular transformation. JSP loading is done with AJAX.

C. Tool configuration

In order to use the ETL tool, administrator has to pre-configure it. Most important are the following parts:

- *Source types* (Fig. 8) – it refers to the source types that the tool can work with (for now PostgreSQL, MySQL and flat file with delimiter)

```
public class Source_MySQL
 implements Extraction_I {
  public static Source_MySql
           get_instance(String args[]){
    Source_MySQL instance =
                    new Source_MySQL();
    return instance;
  }
  public boolean load_parameters(
    String address, String name,
    int port, String username,
    String password){...}
  public Vector get_table_columns(){...}
  public Vector execute_query(
   String query, Vector info){...}
}
```

Fig. 5. Example 1 Example of dynamic loading class

```
public interface Extraction_I {
  public boolean load_parameters(
    String address, String name,
    int port, String username,
    String password);
  public Vector get_table_columns();
  public Vector execute_query(
    String query, Vector info);
}
```

Fig. 6. Example 2 Example of the interface that dynamic loading class must implement

```
public class Extraction {
  private Extraction_I extraction_i;
  public boolean set_class_instance(
   String src_class) {
    Thread t = Thread.currentThread();
    ClassLoader c =
              t.getContextClassLoader();
    Class toRun = null;
    try{toRun = c.loadClass("subsys_ext."
                        +src_class); ...}
    Method mainMethod = null;
    try{mainMethod =
        findMain(toRun,"get_instance");
        }...
        Object instance = null;
        try{ instance =
              mainMethod.invoke(null, new
               Object[]{new String[1]});
            }...
        extraction_i =
                (Extraction_I) instance;
        return true;
  }
  private Method findMain(
  Class my_class, String function_name) {
    Method[] methods =
                  my_class.getMethods();
    for (int i = 0; i < methods.length;
         i++) {
          if (methods[i].getName()
              .equals(function_name))
                 return methods[i];
          }
          return null;
    }}
  public void some_method()
  {...
    extraction_i.load_parmeters(address,
filename, port, username, password);...}}
```

Fig. 7. Example 3 Example of a class that dynamically loads another class [8, p. 11]

- *Transformations* (Fig. 9) – defines which transformations does the tool support, defines the names of classes that implement some particular functionality and the corresponding JSP file which is loaded when the user chooses this transformation.
- *Checkpoints or steps* (Fig. 10) – administrator has to define steps that user follows when filling in the metadata (the administrator must define the page (a JSP file) that opens when user is on a particular step as well as the checkpoints name);

As we mentioned earlier, the program guides the user through the entire process. Fig. 10 shows the steps (checkpoints) for the user; the user has to define how much sources are going to be used.

After that (Fig. 11) we see the input form that is used to define a new data source (there is new PostgreSQL source defined). It is always possible to choose from already existing sources. The tool will use that info and will connect to the source and will retrieve metadata as well. When we have all (sources) metadata and we have defined dimension and fact tables with attributes, the user must define all merges of the attributes (Fig. 12) (for example merge of first and last names into the attribute buff_name_surname).

After this is defined, the user can connect the attributes from the source to destination attributes and define transformations that need to be done. When this step is done for all dimensions/fact tables and the corresponding sources, the last step starts the thread for ETL processing. Before starting the thread the user can change entered data and go back to previous steps.

V. Conclusion

The ETL process is the most important and most problematic part when creating data warehouses. In order to speed up the whole process and in order to make it easier (for users), we built a tool that leads the users through the whole process. Although this ETL tool is far from being perfect and cannot be measured with professional tools on the market, its major advantage is that it is web based; no installation is needed, it is available right away, more users can use it at the same time and users can learn the ETL process when using the tool. The ETL tool is good for users who are not that familiar with the ETL process and who have no time to analyze new ETL tools but want to summarize data, move data into the data warehouse and analyze data. The ETL tool is flexible and because of that it can be easily upgraded.

VI. Future Works

Because this tool is only a prototype, there are many possible improvements. Some parts are already improved; some complex queries were made that extracted more data at once, some filters were implemented to retrieve relevant data (to speed up the tool) etc. In the future we plan to optimize the tool (speed, design, source code, DB queries, security), add new features (add new data sources, new transformations, etc.) and test the tool with larger set of data and compare results to other tools. Also, it is planned to take data from two grocery stores (data from a small data warehouse that was implemented a few years ago) and test the ETL tool with that data and compare it first to manual ETL, and later with other tools. When this is done and tool is optimized, it is planned to do a research with experts where experts should give feedback about usage of the tool in comparison to the tools they are using right now.

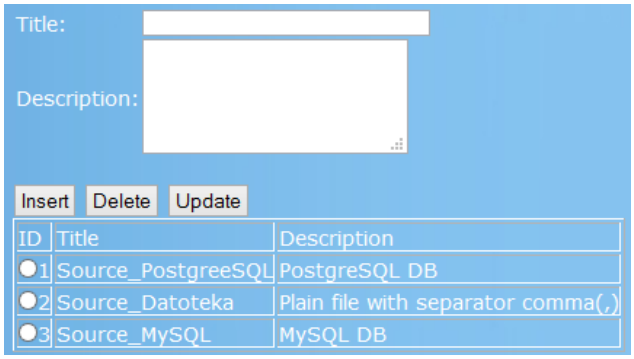


Fig. 8. Administration view of source types

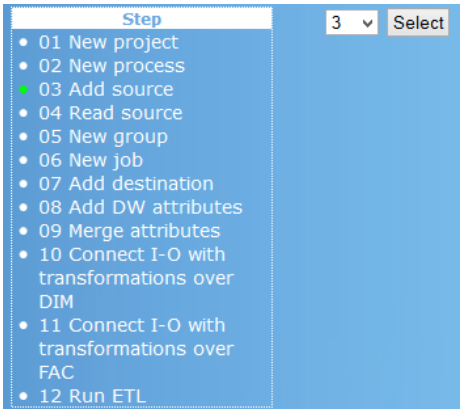


Fig. 9. Menu of checkpoint (steps) for the user (left) and form to select number of sources (right)

Transformation name	Transformation purpose	Physical calculations	Logical calculations	Transformation type
none	no transformation	-	-	none
FIL_format_date	date format change	-	-	FIL_
UN_merge_att_source	merge attributes from source	-	-	UN_
UN_connect_FAC_DIM	connecting fact table with dimensional table	-	-	UN_
FIL_not_null	check if value is null when yes inserts default value	-	-	FIL_
FIL_upper_lover_case	converts all uppercase or lowercase	-	-	FIL_

Screen ID	Transformation name	Screen type	Screen category	ETL stage	Default seventy score	Screen SQL	Exception action
1	none	0	0	STG	-	-	-
2	FIL_format_date	1	1	STG	-	-	-
3	UN_merge_att_source	2	2	STG	-	-	-
4	UN_connect_FAC_DIM	2	2	STG	-	-	-
5	FIL_not_null	2	2	STG	-	-	-
6	FIL_upper_lover_case	2	2	STG	-	-	-

Fig. 10. Administration view of existing transformations and corresponding screen dimension

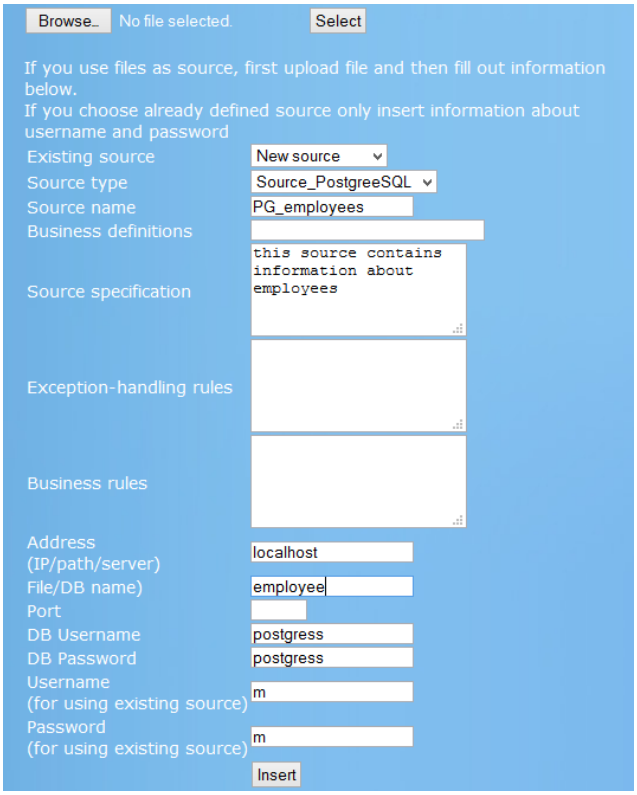


Fig. 11. Form for entering new source

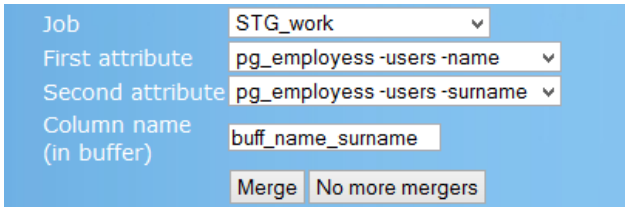


Fig. 12. Form to define attribute merges

REFERENCES

[1] K. Rabuzin and M. Novak, "Data warehouses and ETL," Methods and Tools for Information and Business Systems development (Case22), Zagreb, Jun. 2010, pp. 85-89

[2] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data, Indianapolis: Wiley Publishing Inc., 2004.

[3] C. White: "OLAP in the Database - Intelligent Business Strategies" June 2003. http://www.information-management.com/issues/20030601/6807-1.html?pg=2. [Accessed 3 August 2010].

[4] R. Kimball R., M. Ross, W. Thornthwaite, J. Mund and B. Becker, The Data Warehouse Lifecycle Toolkit – Second Edition, Indianapolis: Wiley Publishing, Inc., 2008.

[5] H. W. Inmon, Building the Data Warehouse – Third Edition, New York: John Wiley & Sons Inc., 2002.

[6] F. Silvers, Building and Maintaining a Data Warehouse, Boca Raton: CRC Press, 2008.

[7] I. M. M. Awad, S. M. Abdullah and M. A. B. Ali, "Extending ETL framework using service oriented architecture", Procedia Computer Science, vol. 3, 2011., pp. 110-114

[8] T. Neward, "Understanding Class.forName - Loading Classes Dynamically from within Extensions" 2000. http://media.techtarget.com/tss/static/articles/content/dm_classForname/DynLoad.pdf. [Accessed 5 July 2010].

[9] P. Ponniah, Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals, New York: John Wiley & Sons Inc., 2001.

Using an MPI Cluster in the Control of a Mobile Robots System

Mohamed Salim LMIMOUNI, Saïd BENAISSA, Hicham MEDROMI, Adil SAYOUTI
Equipe Architectures des Systèmes (EAS), Laboratoire d'Informatique, Systèmes et Energies Renouvelables (LISER)
Hassan II University–Aïn Chock, Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM)
Casablanca, Morocco

***Abstract*—Recently, HPC (High Performance Computing) systems have gone from supercomputers to clusters. The clusters are used in all tasks that require very high computing power such as weather forecasting, climate research, molecular modeling, physical simulations, cryptanalysis, etc. The use of clusters is increasingly important in the scientific community, where the need for high performance computing (HPC) is still growing. In this paper, we propose an improvement of a mobile robots system control by using an MPI (Message Passing Interface) cluster. This cluster will launch, manipulate and process data from multiple robots simultaneously.**

Keywords—clusters; MPI; parallel programming; mobile systems; mobile robots

I. Introduction

Today, almost all industries require fast processing power. With the increasing availability of cheaper and faster computers, many companies are interested in the technological benefits. Supercomputers were developed to meet the needs of fast processing. However, a typical supercomputer can cost usually over a million dollars because of specialized hardware and software. Therefore, clusters have been presented as an alternative to supercomputers.

A cluster is defined as a group of independent computers linked with a computational network and operating as a single computer [1]. In other words, a cluster is a collection of independent and cheap machines, used together as a supercomputer to provide a solution [1].

Using clusters has many advantages:

- The computers that form a cluster are less expensive than supercomputers.
- You can add other nodes (computers) to the cluster as needed.
- On clusters, you can use open source software to reduce software costs.
- Clusters allow multiple computers to work together to solve several problems.

Mobile robots are used as autonomous systems to make autonomous processes more flexible and efficient. In this kind of application, mobile robots systems are guided and controlled by supervisory systems. There are numerous scientific papers about parallel algorithm and scheduling methods in this area but there is no reliable implementation on mobile robots systems in the industry.

By using an MPI cluster in the control of a mobile robots system, we can launch multiple mobile robots at the same time, it also allows us to handle a large number of mobile robots in real time as well as the treatment of recovered data from different sensors of these mobile robots more efficiently.

In this paper, some cluster definitions, a short history, the typical architecture, various types of clusters and MPI are presented in the second section. The third section gives a short view on mobile robots. The fourth section shows the existing architecture and the proposed improvement. In the fifth section, we describe the implementation.

II. Clusters

A. Definitions

A cluster is a single system comprised of interconnected computers that communicate with one another either via a message passing; or by direct, internode memory access using a single address space [2]. We can also define a cluster as a commonly found computing environment consisting of many PCs or workstations connected together by a local-area network [3].

B. History

The official technical basis of clusters as a way to do a parallel work of any kind was probably invented by Gene Amdahl of IBM, who released in 1967 what is considered the seminal paper of parallel processing: the Amdahl's law [4]. Amdahl's Law describes mathematically the performance gain expected after the parallelization of a task on a parallel architecture.

But if we take a glance at the history of clusters, we find that the first commercial cluster was ARCnet [5], developed by Datapoint in 1977. ARCnet was not a commercial success and clusters have really taken off after the development of VAXcluster by Digital Equipment Corporation in 1984 for the VAX / VMS operating system.

The history of clusters would not be complete without mentioning the crucial role played by the development of the Parallel Virtual Machine (PVM) in 1989 [6]. This open source software based on the TCP/IP protocol has allowed the creation of a virtual supercomputer from any system connected through a TCP/IP network.

PVM and the availability of cheap computers have led, in 1993, to a NASA project that consists on the development of supercomputers based on clusters. In 1995, the Beowulf cluster [7] has emerged.

Today, the clusters of Beowulf type are ubiquitous and occupy the top spots in the TOP500 site [8] [9].

C. Typical architecture

The typical architecture of a cluster is shown in Figure 1 [10]. A node of the cluster can be a single or multiprocessor system, such as a PC, workstation, or SMP (Symmetric MultiProcessor). The nodes must be connected via a LAN (Local Area Network) based on Ethernet, Myrinet or InfiniBand. The cluster middleware offers an illusion of a united system of the independent nodes. Parallel programming environments offer portable, efficient and easy-to-use tools for developing parallel applications.

D. Types

There are three varieties of clusters, each one offers different benefits for the user. These varieties are:

- **Load balancing clusters:** used to provide a single interface for a set of resources that can grow arbitrarily. We can imagine a web server that redirects client requests to another node when it has reached its limit of load. This is called "load balancing". Only the node that handles the distribution is visible from the outside.
- **High performance clusters:** They consist of a set of computers linked together to provide maximum power in solving a problem. The heart of these clusters is formed of compute nodes that will receive the code to execute. On smaller clusters we can count ten nodes, while the largest have more than 80 000. The network architecture in place for communication between nodes becomes very heavy, expensive and it limits its performance. You should know also that the ratio of the number of nodes and the performance of this type of Clusters is not linear. It is necessary that the program executed is highly parallelizable and that it requires little communication between the computing units.
- **High availability clusters:** The High Availability clusters are built to provide a secure and fault tolerant environment. The redundancy is the most used method. It consists on multiplying the material that could be subject to failure. Server applications are installed the same way on the cluster nodes.

E. MPI

MPI (Message Passing Interface) is a specification for a standard library for message passing that was defined by the MPI Forum, a broadly based group of parallel computer vendors, library writers, and applications specialists [11]. MPI is not an IEEE or ISO standard, but has become the industry standard for writing message passing programs on HPC platforms.

There are many reasons for using MPI:

- **Standardization:** MPI is the only message passing library which can be considered a standard. It is supported on virtually all HPC platforms. Practically, it has replaced all previous message passing libraries.
- **Portability:** There is little or no need to modify your source code when you port your application to a different platform that supports the MPI standard.
- **Performance Opportunities:** Vendor implementations should be able to exploit native hardware features to optimize performance.
- **Functionality:** There are over 440 routines defined in MPI-3, which includes the majority of those in MPI-2 and MPI-1.
- **Availability:** A variety of implementations are available, both vendor and public domain.

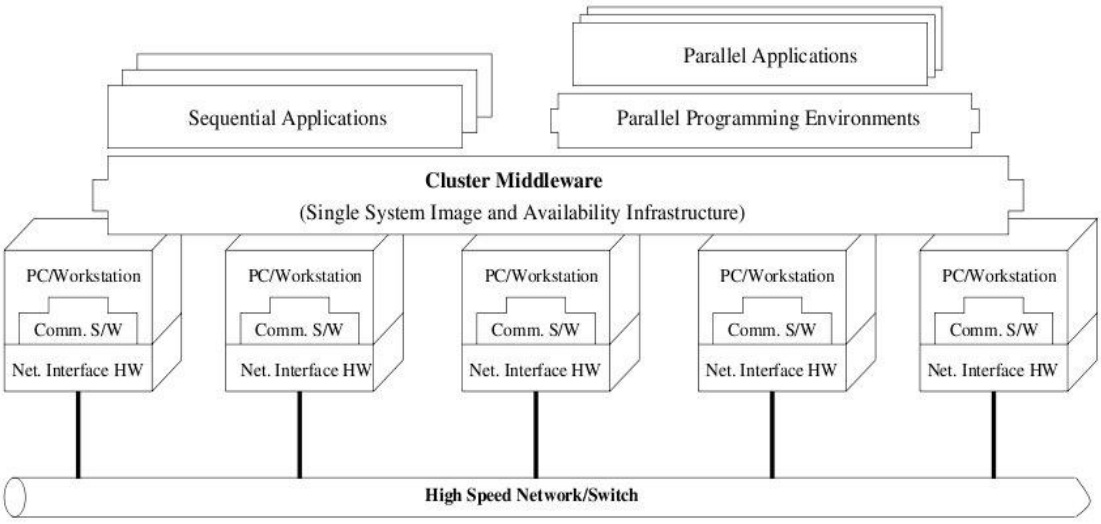


Fig. 1. Typical architecture of a cluster

III. Mobile Robots

A. Definitions

The word robot was introduced in 1921 by Czech writer Karel Capek in his play RUR (Rossum's Universal Robots), as the name for the artificial people created by an unseen inventor (Rossum) to replace humans in any job. Capek described these artificial people as robots, a word derived from the Czech word robota meaning literally serf labor. Robots were constructed to serve and to free humans from any type of labor [12].

This concept has evolved over time and several definitions for the word now exist in the literature.

In mobile robot systems, robot control architecture is necessary. Robot control architecture is defined as a mapping of sensory information into actions in the real world, in order to accomplish a certain task. It is a way of integrating different kinds of hardware and software modules. Furthermore, it plays an important role in maintenance tasks and in the addition of new modules.

B. Khepera III

The K-team [13] robot, Khepera III [14] is a small, circular mobile robot running on two wheels and a sliding support. The diameter is about 130 mm, the height about 70 mm and the weight about 690 g. Different views of the Khepera III are presented in Figure 2, the top row showing a prototype and the bottom row the current commercially available version.

In its basic configuration, the Khepera III is equipped with two motors with associated controllers, a ring of 9 infrared (IR) sensors attached to the bottom layer of the robot's internal structure, another ring of ultrasonic (US) sensors attached to the second layer and an additional pair of IR sensors pointing downward (called ground sensors). Communication with and control of these devices is mediated by a dsPIC 30F5011 microprocessor.

Fig. 2. Khepera III

IV. Proposed Improvement

A. Existing architecture

As part of research conducted within the EAS (Equipe Architectures des Systèmes) team, a mobile system architecture was proposed in 2008 (Figure 3) [15].

The user, through its web browser (Internet Explorer, Google Chrome, etc.), connects to a web server (step 1) using the TCP/IP protocol and download an application on his workstation (step 2). A connection is established to the server responsible for the management of targeted mobile systems (step 3) and after verification, the user can take its control. Parallel to the 3rd step, connections are also established to multimedia servers offering views (video, sound) of the controlled system.

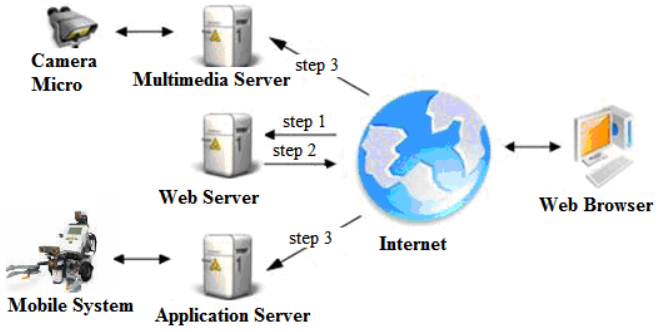


Fig. 3. Mobile system architecture proposed by the EAS team

B. Proposed architecture

In the previous paragraph, we presented the mobile system architecture proposed by the EAS team. This architecture presents some limitations in the Application Server that manages the mobile robots if they are numerous or if they use data from multiple sensors.

To overcome these limitations we propose the architecture in Figure 4 [16] where we replace the Application Server on the old architecture with an MPI Cluster that can launch

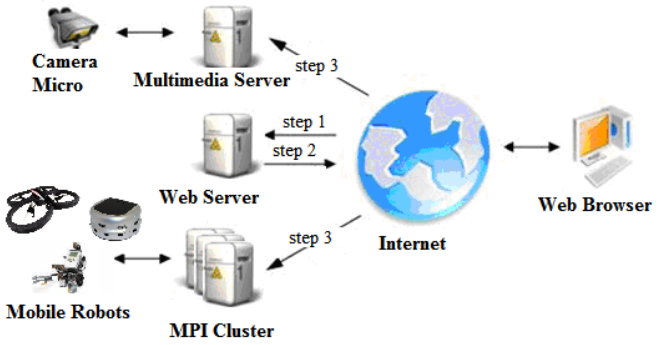


Fig. 4. Improved architecture of the mobile system proposed by the EAS team

multiple mobile robots at the same time, handle data from multiple mobile robots in real time and process recovered data from different sensors of these mobile robots more efficiently.

Figure 5 shows the different layers of the proposed improvement:

- **MMI (Man Machine Interface):** The layer that will allow users to launch, manipulate and process data of mobile robots using graphical tools.
- **Parallel Program:** The code that will launch several mobile robots at the same time, manipulate and process data of these mobile robots in a parallel way.
- **MPI:** The library that will compile and run parallel programs on the cluster.
- **Cluster:** The different nodes (machines) of the cluster. These nodes are connected by a network.
- **Communication:** The layer that will ensure the communication between the cluster's machines and the mobile robots.

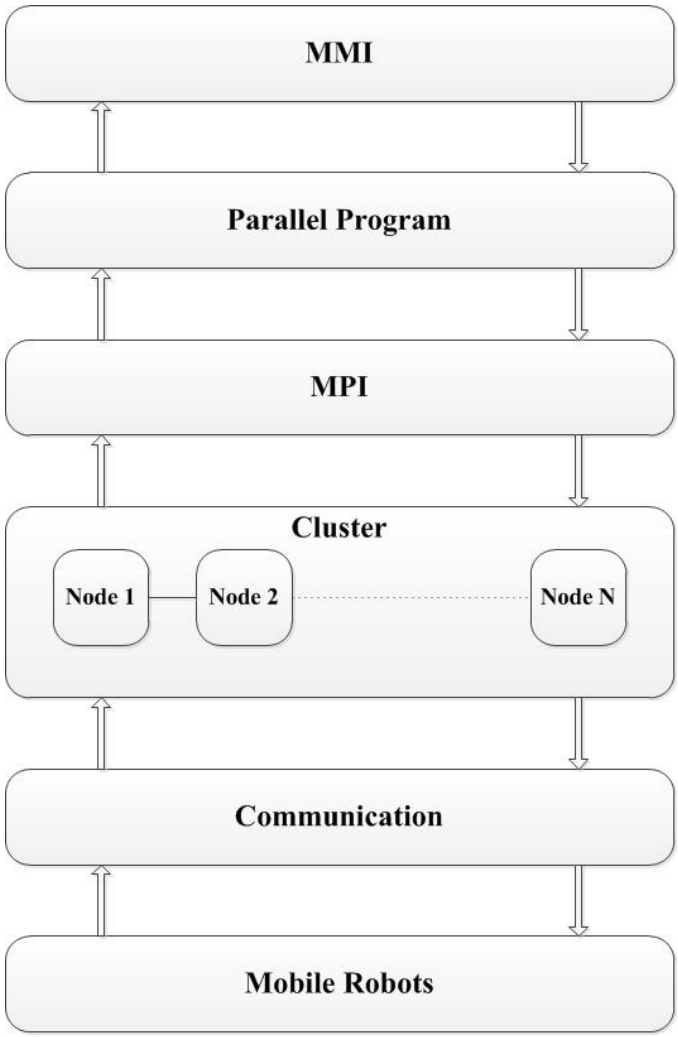


Fig. 5. Layers of the proposed improvement

- **Mobile Robots:** The different mobile robots that will be handled and whose data will be processed using the cluster.

V. Implementation

A. Virtual MPI cluster

We need at least two machines if we want a real cluster. In our case, we built a virtual MPI cluster based on two Ubuntu [17] virtual machines on a DELL OPTIPLEX 780:

- **Processor:** Intel Core 2 Quad Q8400 @ 2.66GHz
- **RAM:** 4 GB
- **HDD:** 500GB

We installed and configured some packages on the two Ubuntu virtual machines:

- **GNU C compiler and GNU FORTRAN compiler:** to be able to compile parallel programs on our virtual MPI cluster.
- **OpenSSH Server and OpenSSH Client:** to use SSH as the way to communicate between different virtual machines.
- **MPICH [18]:** to compile and run parallel programs on our virtual MPI cluster.

Figure 6 shows the components of our virtual MPI cluster.

B. Cluster configuration

We configured our virtual MPI cluster to be able to communicate with the K-team robot, Khepera III.

The challenge posed by the Khepera III platform concerns the setup of a cross compilation tool chain for the ARM Linux system running on the KoreBot board, i.e. an environment that allows to compile executable programs (e.g. from C or C++ source code) for the ARM Linux system.

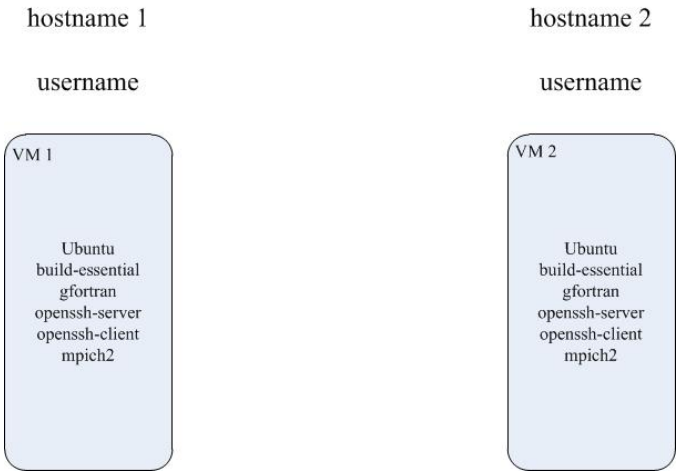


Fig. 6. Virtual MPI cluster

C. Parallel algorithm

We used C language and MPI to program the parallel launch of two Khepera III robots.

The program is executed on our virtual MPI cluster. Each node is responsible of the launch of a robot.

The algorithm pseudo code is demonstrated as follows:

```
BEGIN
        Initialize MPI environment
        IF node1 THEN
                Launch robot1
        ELSE
                Launch robot2
        ENDIF
        Terminate MPI environment
END
```

VI. CONCLUSIONS

The growing need for powerful and cheaper computer processors in the world increases the use of clusters. Today, clusters are used in several areas (commercial, scientific, etc.). The field of mobile robots system, an area in large changes, also needs to exploit the advantages of the use of clusters.

That is why we have proposed in this paper the use of an MPI cluster in the control of a mobile robots system to improve its performances and overcome its limitations.

Creating graphical interfaces for interacting with users and testing the parallel processing of data sent by multiple mobile robots are among the perspectives of our work.

REFERENCES

[1] S. Aydin and O. F. Bay, "Building a high performance computing clusters to use in computing course applications" Procedia - Social and Behavioral Sciences, vol. 1, pp. 2396-2401, Feb. 2009.

[2] G. Bell and J. Gray, "What's Next in High-Performance Computing", Communications of the ACM, vol. 45, issue 2, pp. 91–95, Feb. 2002.

[3] J. Dongarra, "Trends in high performance computing: a historical overview and examination of future developments", Circuits and Devices Magazine, IEEE, vol. 22, Issue 1, pp. 22-27, Feb. 2006.

[4] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities", in *Proc. AFIPS'67*, 1967, p. 483-485.

[5] ARCNET Resource Center. [Online]. Available: http://www.arcnet.com

[6] PVM: Parallel Virtual Machine. [Online]. Available: http://www.csm.ornl.gov/pvm/

[7] Beowulf.org | Mailing List and Archives. [Online]. Available: http://www.beowulf.org/

[8] Home | TOP500 Supercomputer Sites. [Online]. Available: http://www.top500.org/

[9] V. Kindratenko and P. Trancoso, "Trends in High-Performance Computing", Computing in Sciences & Engineering, vol. 13, Issue 3, pp. 92-95, Jun. 2011.

[10] R. Buyya, "*High Performance Cluster Computing: Programming and Applications"*, Prentice Hall PTR, NJ, USA, 1999.

[11] W. Gropp, E. Lusk, N. Doss and A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard" Elsevier Parallel Computing, vol. 22, Issue 6, pp. 789-828, Sept. 1996.

[12] R. C. Arkin, *Behavior-Based Robotics*, The MIT Press, 1998.

[13] K-Team Corporation | Mobile Robotics. [Online]. Available: http://www.k-team.com/

[14] Khepera III. [Online]. Available: http://www.k-team.com/mobile-robotics-products/khepera-iii

[15] A. Sayouti, "Conception et Réalisation d'une Architecture de Contrôle à Distance Via Internet à Base des Systèmes Multi-Agents", PhD thesis, Ecole Nationale Supérieure d'Electricité et de Mécanique, Hassan II University, Casablanca, Morocco, Jul. 2009.

[16] M. S. Lmimouni, H. Medromi and A. Sayouti, "Utilisation d'un "cluster" HPC dans le contrôle via internet d'un système mobile", in Proc. JDTIC'12, 2012.

[17] The world's most popular free OS | Ubuntu. [Online]. Available: http://www.ubuntu.com/

[18] MPICH | High-Performance Portable MPI. [Online]. Available: http://www.mpich.org/

Simulation of Performance Execution Procedure to Improve Seamless Vertical Handover in Heterogeneous Networks

Omar Khattab and Omar Alani
School of Computing, Science & Engineering
University of Salford
Greater Manchester- M5 4WT,UK

***Abstract*— One challenge of wireless networks integration is the ubiquitous wireless access abilities which provide the seamless handover for any moving communication device between different types of technologies (3GPP and non-3GPP) such as Global System for Mobile Communication (GSM), Wireless Fidelity (Wi-Fi), Worldwide Interoperability for Microwave Access (WiMAX), Universal Mobile Telecommunications System (UMTS) and Long Term Evolution (LTE). This challenge is important as Mobile Users (MUs) are becoming increasingly demanding for services regardless of technological complexities associated with it. To fulfill these requirements for seamless Vertical Handover (VHO) two main interworking architectures have been proposed by European Telecommunication Standards Institute (ETSI) for integration between different types of technologies; namely, loose and tight coupling. On the other hand, Media Independent Handover IEEE 802.21 (MIH) is a framework which has been proposed by IEEE Group to provide seamless VHO between the aforementioned technologies by utilizing these interworking architectures to facilitate and complement their works. The paper presents the design and the simulation of a Mobile IPv4 (MIPv4) based procedure for loose coupling architecture with MIH to optimize performance in heterogeneous wireless networks. The simulation results show that the proposed procedure provides seamless VHO with minimal latency and zero packet loss ratio.**

Keywords—Vertical Handover (VHO); Media Independent Handover (MIH); Interworking Architectures; Mobile IPv4 (MIPv4); Heterogeneous Wireless Networks

I. INTRODUCTION

With the advancement of wireless communication and computer technologies, mobile communication has been providing more versatile, portable and affordable networks services than ever. Therefore, the number of Mobile Users (MUs) communication networks has increased rapidly. For example, it has been reported that "today, there are billions of mobile phone subscribers, close to five billion people with access to television and tens of millions of new internet users every year" [1] and there is a growing demand for services over broadband wireless networks due to diversity of services which can't be provided with a single wireless network anywhere anytime [2]. This fact means that heterogeneous environment of wireless networks, such as Global System for

Mobile Communication (GSM), Wireless Fidelity (Wi-Fi), Worldwide Interoperability for Microwave Access (WiMAX) and Universal Mobile Telecommunications System (UMTS) will coexist providing MU with roaming capability across different networks. One of the challenging issues in Next Generation Wireless Systems (NGWS) is achieving seamless Vertical Handover (VHO) while roaming between these technologies. Therefore, telecommunication operators will be required to develop a strategy for interoperability of these different types of existing networks to get the best connection anywhere anytime. To fulfill these requirements of seamless VHO two main interworking architectures have been proposed by European Telecommunication Standards Institute (ETSI); namely, loose and tight coupling for integration between the different types of technologies. On the other hand, Media Independent Handover IEEE 802.21 (MIH) is a framework which has been proposed by IEEE Group to provide seamless VHO between different technologies by utilizing the above interworking architectures to complement their works. The paper presents the design and the simulation of a MIPv4 based procedure for loose coupling architecture with MIH to optimize performance in heterogeneous wireless networks in terms of latency and packet loss. The results of the proposed procedure show that it can provide a seamless VHO with minimal latency and zero packet loss ratio.

The rest of the paper is organized as follows: section II describes the VHO management. In section III, related works are presented. In section IV, the proposed procedure is presented. In section V, the simulation results and discussions of the proposed procedure are presented and finally, the conclusion is included in section VI.

II. VERTICAL HANDOVER MANAGEMENT

The mechanism which allows the MUs to continue their ongoing sessions when moving within the same Radio Access Technology (RAT) coverage areas or traversing different RATs is named Horizontal Handover (HHO) and VHO, respectively. In the literature most of the research papers have divided the VHO management into three phases: Collecting Information, Decision and Execution (e.g., [3 and 4]) as described below.

Handover Collecting Information

In this phase, all required information for handover decision is gathered, some of this information is related to the user's preferences (e.g., cost, security), network (e.g., latency, coverage) and terminal (e.g., battery, velocity).

Handover Decision

In this phase, the best RAT based on aforementioned information is selected and the handover execution phase is informed about that.

Handover Execution

In this phase, the active session for the MU will be maintained and continued on the new RAT. After that, the resources of the old RAT are eventually released.

III. Related Works

In previous works, three surveys about VHO approaches proposed have been presented [5, 6 and 7].

In [5], the VHO approaches proposed in the literature have been classified into four categories based on MIH and IP Multimedia Subsystem (IMS) frameworks (MIH based VHO category, IMS based VHO category, MIP under IMS based VHO category and, MIH and IMS combination based VHO category) in order to present their objectives in providing seamless VHO. It has been concluded in [5] that the MIH is more flexible and has better performance providing seamless VHO compared with IMS framework. The IEEE Group has proposed MIH to provide a seamless VHO between different RATs [8 and 9]. The MIH defines two entities: first, Point of Service (PoS) which is responsible for establishing communication between the network and the MU under MIH and second, Point of Attachment (PoA) which is the RAT access point. Also, the MIH provides three main services: Media Independent Event Service (MIES), Media Independent Command Service (MICS) and Media Independent Information Service (MIIS) [10] such that the MIH relies on the presence of mobility management protocols (e.g., MIPv4 and MIPv6).

In [6], the VHO approaches proposed in the literature have been classified into two categories based on the mobility management protocols (MIPv4 and MIPv6) for which their performances and characteristics have been presented. It has been concluded in [6] that providing service continuity through MIPv4 category under MIH will allow the operators to diversify their access networks take into account advantages of this category while MIPv6 category under MIH requires future work improvements in terms of VHO decision criteria, additional entities, complexity, diversity of RATs and evaluation using empirical work real environment.

In [7], loose and tight coupling interworking architectures have been surveyed their objectives, features and challenges. It has been concluded in [7] that the loose coupling is more suitable with MIH and contributes for enhancing its vital role in heterogeneous wireless environment to get fast and soft seamless roaming with minimal latency and minimal packet loss, respectively. From previous works [5, 6 and 7], three vital things have been concluded as follows:

- MIH is more flexible and has better performance providing seamless VHO compared with IMS framework.
- MIPv4 category under MIH will allow the operators to diversify their access networks take into account advantages of this category while MIPv6 category under MIH requires future work improvements.
- Loose coupling is more suitable with MIH and contributes for enhancing its vital role in heterogeneous wireless environment (minimal latency and minimal packet loss).

As a result of the conclusions above, a procedure of loose coupling which could be applied in conjunction with MIPv4 under MIH has been proposed in [11 and 12]. In [12], analytical modeling results considering Wi-Fi and WiMAX scenario showed that the VHO latency and packet loss were significantly reduced compared with the procedures found in the literature: Proxy MIPv6 (PMIPv6), Proxy First MIPv6 (PFMIPv6) and MIH-enabled PMIPv6. The results in [12] showed that the proposed procedure outperformed the existing procedures and scored ($4.4x10^{-3}$ sec) and ($1.6x10^{-2}$) of latency and packet loss ratio, respectively.

IV. The Proposed Procedure

This section describes the proposed procedure through VHO phases: Initiation, Decision and Execution.

A. Initiation Phase

In this phase, while MU is connected to a source network the VHO procedure will be triggered imperatively due to Radio Signal Strength (RSS) going down or alternatively based on the user's preferences (e.g., high data rate, low cost).

B. Decsion Phase

In this phase, as a result of triggering in the initiation phase, *MIIS Request/Response Available RATs* message will be responsible to pass available RATs to MU via source network (PoA and PoS). In imperative session due to RSS going down the MU will select RATs list of priority based on user's preferences and then pass them to the destination PoS via source network whereas in alternative session the MU will select RATs list of priority based on user's preferences due to his/her profile change.

When the first choice from RATs list of priority could not be satisfied with available resources, the Admission Control (AC) at destination PoS will automatically move to another RAT selection in the list in order to satisfy the requirements of this RAT selection and so on. Once RAT of sufficient resources has been found, it will be checked by destination PoS whether it is compliant with the rules and preferences of operators. If that is available, the MIIS/Home Agent (HA) will be informed to start buffering for new data packets which are sent by Correspondent Node (CN).

C. Execution Phase

In this phase, the MU will be connected to target RAT to start its Authentication, Authorization and Accounting (AAA) with destination PoA and obtain Care of Address (CoA) from Dynamic Host Configuration Protocol (DHCP).

After that, *Update/Acknowledge binding* message notifies HA about the new CoA to start sending the buffered data and continuing the session within target RAT, this is shown in Fig. 1.

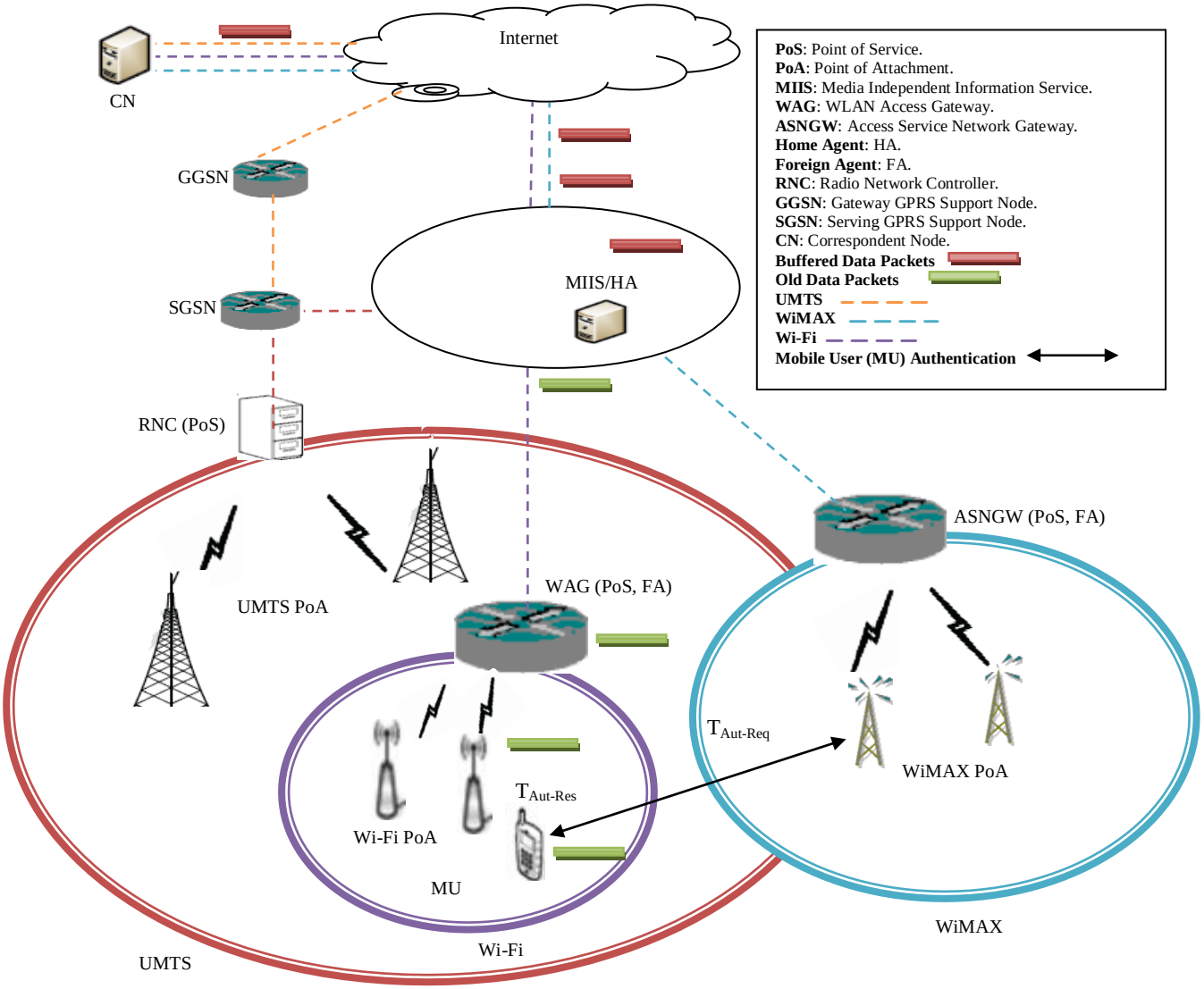


Fig. 1. Diagram of proposed procedure [12]

V. Simulation Results and Discussions

The latency and packet loss are the major drawbacks in the execution phase where this phase is out of the scope of MIH (e.g., handover signaling, context transfer and packet reception) [13]. Therefore, after the analysis of the results in previous work [12], the simulation has applied the proposed procedure of loose coupling in conjunction with MIPv4 taking into account the handover signaling time in the execution phase and RSS going down in order to make VHO decision. The MU originally is hosted by Wi-Fi and it has started moved toward the WiMAX and received VoIP traffic, this is shown in Fig. 2. Detailed characteristics of the simulation parameters are explained in TABLE I.

After the implementation of the proposed procedure in the specific scenario, Fig. 3 and Fig. 4 illustrate the proposed procedure with average latency of ($2x10^{-5}$ sec) and zero packet loss, respectively. The latency is the time taken for the MU to obtain a new IP address from a target network and register itself with HA [14]. During this process the MU does not receive any packets as a result of handover. The latency is the main cause of packet losses during handover [15]. Therefore, the results obtained in this simulation and the analytical modeling in previous work [12] show that the packet loss ratio improves as long as the latency reduced.

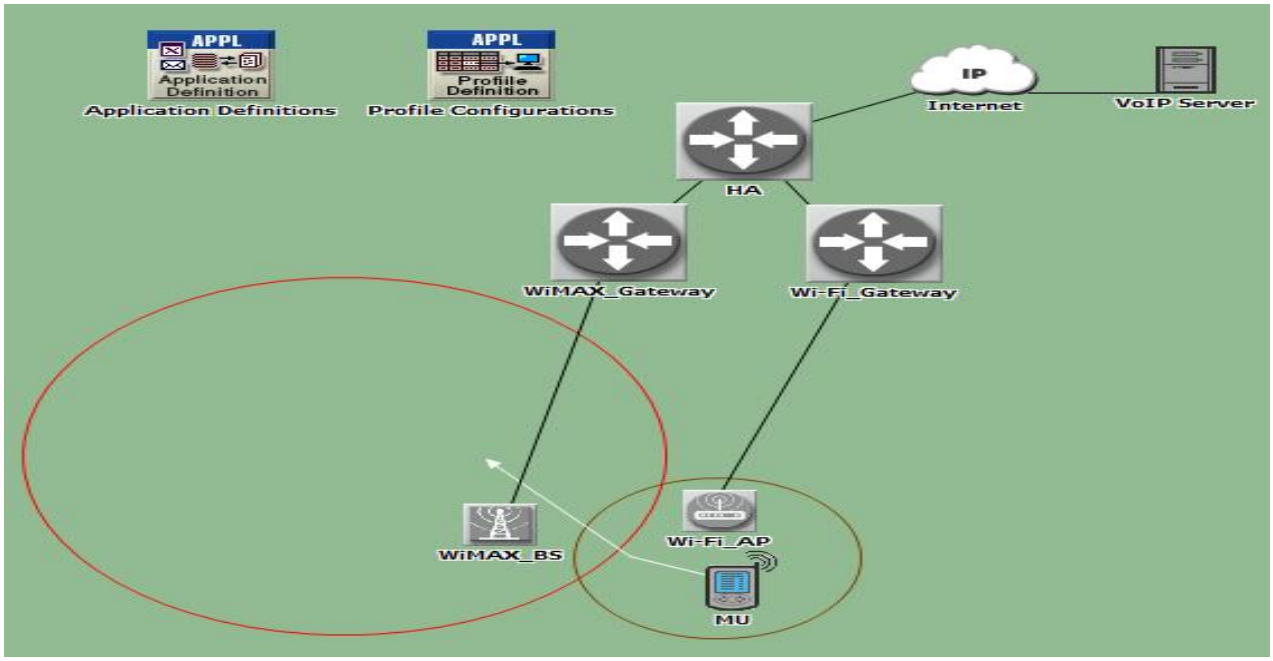


Fig. 2. Simulation diagram of proposed procedure from Wi-Fi to WiMAX

TABLE I. PARAMETERS FOR PERFORMANCE EVALUATION OF SIMULATION MODELING

Name of the Parameter	Value of the Parameter
Simulation Duration	60 minute.
Path (Trajectory)	Linear.
Mobile User Velocity	10 Km/hr.
Traffic	VoIP.
WiMAX	
Cell Coverage	Ellipse, width=1000 m, height=1000 m.
Maximum Transmission Power	0.1 W.
Physical Profile Type	OFDM.
Receiver Sensitivity	-200dBm.
Antenna Gain	15 dBi.
Wi-Fi	
Cell Coverage	Ellipse, width=450 m, height=450 m.
Transmit Power	0.0005 W.
Physical Profile	Direct sequence.
Packet Reception-Power Threshold	-95 dBm.
Data Rate	11 Mbps.

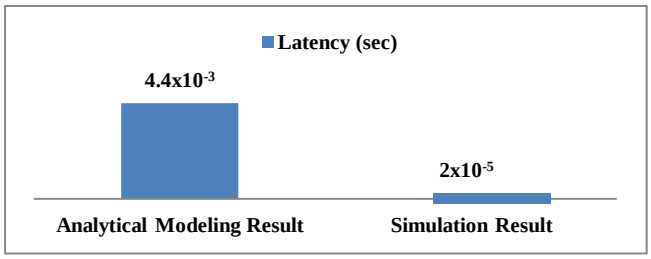


Fig. 3. Comparison of the proposed procedure performance using simulation result vs. analytical modeling result (latency)

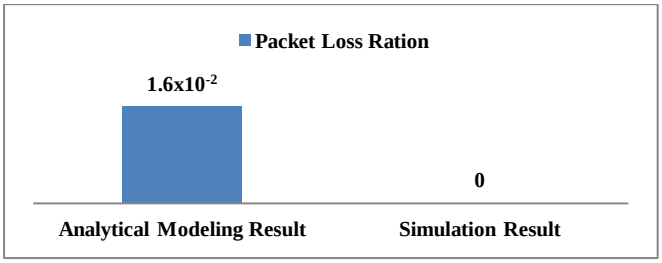


Fig. 4. Comparison of the proposed procedure performance using simulation result vs. analytical modeling result (packet loss)

VI. CONCLUSION

The paper has presented the design and the simulation of a MIPv4 based loose coupling architecture with MIH for providing optimized performance in heterogeneous wireless networks in terms of latency and packet loss.

The simulation results of the proposed procedure show that the VHO latency and packet loss are significantly reduced. In future work, a much more sophisticated and intrinsic scenarios are required which would take into account a wider array of parameters and MUs to make a more intelligent and optimized network selection.

REFERENCES

[1] Our vision: Committed to connecting the world. (2013). International Telecommunication Union (ITU). Retrieved 20 May, 2014, from http://www.itu.int/en/about/Pages/vision.aspx.

[2] B. Angoma, M. Erradi, Y. Benkaouz, A. Berqia, and M.C. Akalay, "HaVe-2W3G: A vertical handoff solution between WLAN, WiMAX and 3G networks," 7th International Wireless Communications and Mobile Computing Conference (IWCMC), pp.101-106, 4-8 July 2011.

[3] M. Louta, P. Zournatzis, S. Kraounakis, P. Sarigiannidis, and I. Demetropoulos, "Towards realization of the ABC vision: a comparative survey of access network selection," Symposium on Computers and Communications(ISCC), pp.472-477, 28 June 2011-1 July 2011.

[4] M. Zekri, B. Jouaber, and D. Zeghlache, "Context aware vertical handover decision making in heterogeneous wireless networks," IEEE 35th Conference on Local Computer Networks (LCN), pp.764-768, 10-14 October 2010.

[5] O. Khattab, and O. Alani, "A survey on Media Independent Handover (MIH) and IP Multimedia Subsystem (IMS) in heterogeneous wireless networks," International Journal of Wireless Information Networks, vol.20, no.2, pp.215-228, July 2013.

[6] O. Khattab, and O. Alani, "Survey on Media Independent Handover (MIH) approaches in heterogeneous wireless networks," 19th European Wireless 2013 (EW 2013), pp.1-5, 16-18 April 2013.

[7] O. Khattab, and O. Alani, "An overview of interworking architectures in heterogeneous wireless networks: objectives, features and challenges," 10th International Network Conference 2014 (INC 2014), 8-10 July 2014. Accepted 31 March 2014 and to be published after the Conference.

[8] P. Neves, J. Soares, and S. Sargento, "Media independent handovers: LAN, MAN and WAN scenarios," IEEE GLOBECOM Workshops, pp.1-6, 30 November 2009 –4 December 2009.

[9] G. Lampropoulos, A.K. Salkintzis, and N. Passas, "Media independent handover for seamless service provision in heterogeneous networks," IEEE Communication Magazine, vol.46, no.1, pp.64-71, January 2008.

[10] J. Marquez-Barja, C.T. Calafate, J.C. Cano, and P. Manzoni, "Evaluation of a technology-aware vertical handover algorithm based on the IEEE 802.21 standard," IEEE Wireless Communications and Networking Conference (WCNC), pp.617-622, 28-31March 2011.

[11] O. Khattab, and O. Alani, "I AM 4 VHO: new apprach to improve seamless vertical handover in heterogeneous wireless networks," International Journal of Computer Networks & Communications (IJCNC), vol.5, no.3, pp.53-63, May 2013.

[12] O. Khattab, and O. Alani, "Mobile IPv4 based procedure for loose coupling architecture to optimize performance in heterogeneous wireless networks," International Journal of Computer Networks and Wireless Communications (IJCNWC), vol.3, no.1, pp.56-61, February 2013.

[13] IEEE 802.21 Tutorial. (2006). IEEE 802.21. Retrieved 20 May, 2014, from http://www.ieee802.org/21/.

[14] S. Haseeb, and A.F. Ismail, "Handoff latency analysis of mobile IPv6 protocol variations," Computer Communications, vol.30, no.4, pp.849-855, 26 February 2007.

[15] Z. Liyan, Z. Li Jun and P. Samuel, "Performance analysis of seamless handover in mobile IPv6-based cellular network," In cellular networks - positioning, performance analysis and reliability, M. Agassi, Ed. Croatia: InTech, 2011, pp.305-330.

Toward an Effective Information Security Risk Management of Universities' Information Systems Using Multi Agent Systems, Itil, Iso 27002,Iso 27005

S.FARIS
EAS Team, LISER Laboratory,
ENSEM
Casablanca, MOROCCO

S.EL HASNAOUI
EAS Team, LISER Laboratory,
ENSEM
Casablanca, MOROCCO

H.MEDROMI
EAS Team, LISER Laboratory,
ENSEM
Casablanca, MOROCCO

H.IGUER
EAS Team, LISER Laboratory,
ENSEM
Casablanca, MOROCCO

A.SAYOUTI
EAS Team, LISER Laboratory,
ENSEM
Casablanca, MOROCCO

***Abstract*—Universities in the public and private sectors depend on information technology and information systems to successfully carry out their missions and business functions. Information systems are subject to serious threats that can have adverse effects on organizational operations and assets, and individuals by exploiting both known and unknown vulnerabilities to compromise the confidentiality, integrity, or availability of the information being processes, stored or transmitted by those systems. Threats to information systems can include purposeful attacks, environmental disruptions, and human/machine errors, and can result in harm to the integrity of data. Therefore, it is imperative that all the actors at all levels in a university information system understand their responsibilities and are held accountable for managing information security risk- that is the risk associated with the operation and use of information systems that support the missions and business functions of their university.**

The purpose of this paper is to propose an information security toolkit namely URMIS (University Risk Management Information System) based on multi agent systems and integrating with existing information security frameworks and standards, to enhance the security of universities information systems.

Keywords—Information security; information systems; multi agent systems; ITIL V3; ISO 27002; ISO 27005

I. INTRODUCTION

Information systems (ISs) are everywhere. They have a large impact on the everyday lives of universities as well as on individuals. At the heart of information systems, security aspects play a vital role and are thus becoming central issues in those systems' effective usage.

The importance of security technologies and of their enabling technical platforms has been widely recognized and receives continuous attention (e.g., new encryption, algorithms, public key infrastructures, etc.).

For some people, security management issues start with updating an antivirus database, but from a more serious perspective, universities understand that security concerns are the source of important costs, not only in terms of technologies but especially in terms of related management activities.

There are emerging calls for an integrated view of information security, from the technological, human, and organizational aspects, sometimes referred as MTO (Man, Technology, and Organization).

However, there is a lack in the methods for tackling the MTO issues in information security. One of the research focuses on the development of information security checklist and standards in order to capture the best practice.

Another research focuses on risk assessment by identifying the threats and vulnerabilities, and then determining the likelihood and impact for each risk. Risk assessment could either be qualitative, categorizing low, medium and high risks, or be quantitative, calculating the value of "Annualized Loss Expectancy"

This paper is presented as follows: after a brief introduction, in section two; a survey of available information security risk management methods and tools will be presented, and then the standards, ISO 27002, ISO 27005, and the framework ITIL will be described. Then, in the third section the toolkit URMIS will be proposed and the multi agent system

will be introduced. The fourth section will propose the architecture, before concluding this paper.

II. State of The Art

A. Risk Management tools and frameworks

An organizational risk is the risk to the organization or to individuals associated with the operation of an information system. The management of organizational risk is a key element in the organization's information security program and provides an effective framework for selecting the appropriate security controls for an information system---the security controls necessary to protect individuals and the operations and assets of the organization.

The common view a Risk Assessment Framework provides helps an organization see which of its systems are at low risk for abuse or attack and which are at high risk. The data an RAF provides is useful for addressing potential threats pro-actively, planning budgets and creating a culture in which the value of data is understood and appreciated.

There are several risk assessment frameworks and risk management methods that are accepted as industry standards that we can list in the figure below.

Attributes / Methods	Risk identification	Risk Analysis	Risk Evaluation	Risk assessment	Risk treatment	Risk acceptance	Risk communication	Languages	Price (method only) (Information assessed in June 2006)	Size of organization	Skills needed[8]	Licensing	Certification	Dedicated support tools
Austrian IT Security Handbook	••	•	•	•••	•••	•••	•••	DE	Free	All	**	N	N	Prototype (free of charge)
Cramm	•••	•••	•••					EN, NL, CZ	Not free	Gov, Large	***	N	N	CRAMM expert, CRAMM express
Dutch A&K analysis	•••	•••	•••					NL	Free	All	*	N	N	
Ebios	•••	•••	•••	•••	•••	•••	•••	EN, FR, DE, ES	Free	All	**	Y	N	EBIOS version 2 (open source)
ISF methods	•••	•••	•••	•••	•••	•••	•••	EN	For ISF members	All except SME	* to ***	N	N	Various ISF tools (for members)
ISO/IEC IS 13335-2 (ISO/IEC IS 27005)	••	••	••	••	•••	•••	•••	EN	Ca. €100	All	**	N	N	
ISO/IEC IS 17799	•				•			EN	Ca. €130	All	**	N	Y	Many
ISO/IEC IS 27001					•	•		EN, FR	Ca. €80	Gov, Large	**	Y	Y	Many
IT-Grundschutz	•••	•••	•••	•••	•••	•••	•••	EN, DE	Free	All	**	Y	Y	Many
Marion (replaced by Mehari)	•••	•••	•••					EN, FR	Not free	Large	*	N	N	
Mehari	•••	•••	•••					EN, FR	€100-500	All	**	N	N	RISICARE
Octave	••	••	••	••	••	••	••	EN	Free	SME	**	N	N	
SP800-30 (NIST)	•••	•••		•••	•••	•••		EN	Free	All	**	N	N	

Fig. 1. Risk Management methods and frameworks

None of these tools implement the multi agent system approach.

Incorporation of the use of information and communication technology in Moroccan universities, involves the need to secure data in information systems.

There is a very little research related to the applications of multi agent systems (MAS) in information system security.

Besides to that, these tools are difficult to use because they require a certain level of knowledge.

Moreover, they don't provide recommendations or immediate solutions to security issues; they just give guidelines to follow in order to ensure an effective security of the information system.

Based on the methodologies aforementioned, and other works described in [4] [5] [6] [10] [11], we propose an integration of the use of ISO 27002, ISO 27005, ITI, and multi-agent systems to develop an information security risk management tool of universities information systems named URMIS (Universities Risk Management Information System).

B. ISO 27002

The ISO 27002 standard is a collection of information security guidelines that are intended to help an organization implement, maintain, and improve its information security management. It is a code of good practices that provides hundreds of potential controls that are designed to be implemented with guidance provided within ISO 27001.

The strengths of ISO 2700 are listed below:

- Optimize the costs of ISS by associating with ISO 27001
- Increased knowledge of risk management
- Does not require a technical solution

Whereas its weaknesses are listed below:

- Optimize the costs of ISS by associating with ISO 27001
- Increased knowledge of risk management
- Does not require a technical solution

In the current version published 2013, ISO 27002:2013 contains 114 controls, as opposed to the 133 documented within the 2005 version. However for additional granularity, these are presented in fourteen sections, rather than the original eleven.

C. ISO 27005

ISO 27005 is intended to provide guidelines for information security risk management. It is used either autonomously or as a support for ISO 27001. It supports the general concepts specified in ISO 27001 and is designed to assist the satisfactory implementation of information security based on a risk management approach. It does not specify or recommend any specific risk analysis method, although it specifies a structured, systematic and rigorous process from analyzing risks to creating the risk treatment plan.

The strengths of ISO 27005 are as follows:

- Flexible and reusable
- Continuous risk management
- Highlighting the human factor: the concept of responsibility
- Whereas its weaknesses are as follows:
- No specific methodology for risk management

The figure below gives an overview of the information security risk management process in ISO 27005.

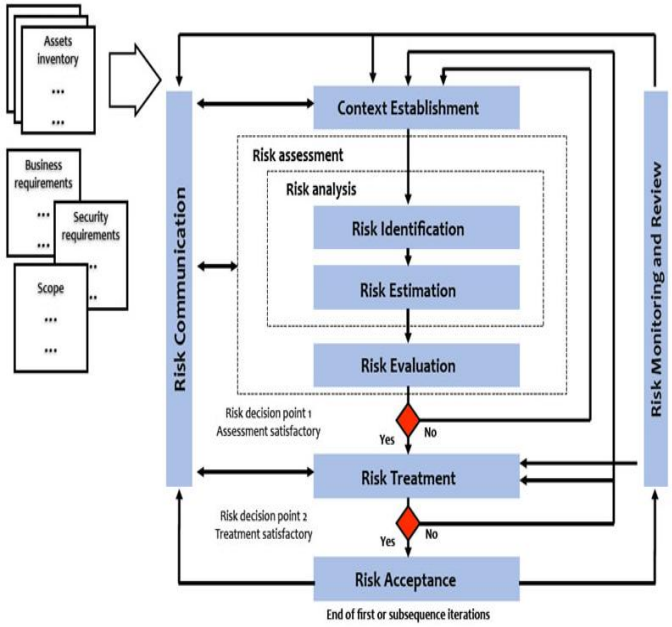


Fig. 2. Information Security Risk Management process

D. ITIL V3

The Information Technology Infrastructure Library (ITIL) is a framework of best practices that promote quality computing services in IT sector.

ITIL is the most widely accepted approach to IT service management in the world. ITIL provides a cohesive set of best practice, drawn from the public and private sectors internationally.

ITIL presents a broad set of management procedures, which apply to all aspects of IT infrastructure, with which an organization can manage its IT operations (Zegers, 2006, Wegmann, 2008).

The ITIL v3 Core consists of five publications, each providing

guidance on a specific phase in the service management lifecycle.

The ITIL Core publications are as follows:

- **Service strategy**
- **Service design**
- **Service transition**
- **Service operation**
- **Continual service improvement**

ITIL can help companies assess their risks, and put procedures in place to log and respond to incidents. ITIL, and more specifically the ITIL security management process, is widely used for the implementation of information security within an organization. ITIL v3 has placed the information security management process within the Service Design core practice book. The goal of the information security management process is to align IT security with business security and ensure that information security is effectively managed in all services and service management activities (OGC, 2007; Taylor,2008).

E. Information security

Confidentiality, integrity and availability are basic requirements for business information security and provide the maintenance requirements of the business (ITGI, 2009), (Kwok and Longley, 1999), (Fitzgerald, 2007), (Sêmola, 2003), (Dias, 2000), (Moreira, 2001).

- **Confidentiality (C)**: All information must be protected according to the degree of privacy of their content, aimed at limiting its access and used only by the people for whom they are intended;
- **Integrity (I)**: All information must be kept in the same condition in which it was released by its owners, in order to protect it from tampering, whether intentional or accidental;
- **Availability (A)**: All the information generated or acquired by an individual or institution should be available to their users at the time they need them for any purpose;

III. Objectives and Importance of The Research

The major objective of this work is to design and implement an integrated toolkit for improving risk management of a university information system.

This work explores how to promote integration and the establishment of a toolkit that would allow each university to have reliable data on higher education, driving better management and improve their governance and risk management.

Implementing this toolkit involves taking a proactive, strategic and measured approach that is more efficient than the reactive one used in many universities. This can be reached across a strategic integration of appropriate frameworks, models and methods in governance and information security.

Analyzing the relevant frameworks, models and methods, used in the aforementioned domains, and extracting the best practices for implementing them in URMIS, can provide effective security of university IS assets.

A. The proposed toolkit

URMIS (Universities Risk Management Information System) is an information security toolkit that provides guidance policies to achieve an effective information security risk management in universities' information systems.

With the intention of implementing the task of information security risk management, URMIS needs to collect data about the status of information asset, recognize kinds of risk, and perform risk management task based on a good defined risk management process. That means the working environment of URMIS consists of knowledge, data, process and strategies. However, knowledge, data, process and strategies are resources in different formalization, and it is a complex work to design interface for each resource. This work is based on the multi agent systems approach, because of its benefits. It encompasses cooperation, resolution of complex problems, modularity, efficiency, reliability and reusability. All these advantages provided by MAS fit these needs.

B. *Agent and Multi agent systems (MAS)*

Jennings and Wooldridge [Jennings & Wooldridge 1998] have defined an agent as "a computer system located in certain environment which is able to act autonomously in this environment, in order to meet its design goals". Agents have the following main properties and characteristics:

- **Autonomy** : agents encapsulate a state (which is not available to other agents), and make decisions on what to do based on this state, without direct human intervention or other persons;
- **Social ability**: agents interact with other agents (and possibly humans) via some kind of agent communication
- Language, and generally have the opportunity to participate in social activities (such as cooperation for solving problems or negotiating) to achieve their goals.
- **Reactivity**: agents are put in an environment (which may be the physical world, a user via a graphical interface, a collection of other agents, the internet, or perhaps many of these combinations), are able to perceive this environment (through the use of potentially imperfect sensors), and are able to respond to timely changes that occur in it.
- **Proactivity**: Agents do not simply act in response to their environment; they are able to solve a problem by taking the initiative.

A multi-agent system (MAS) is a system composed of several intelligent agents that interact with each other. They can be used to solve problems that are difficult or impossible to solve for an individual agent or monolithic system. Multi-agent systems are open and scalable systems that enable the implementation of autonomous and proactive software components. They are characterized by the local autonomy, social interaction, adaptability, robustness and scalability, and for these reasons, they are a very promising paradigm to address the challenges facing automation and check systems.

IV. URMIS INFORMATION SECURITY ARCHITECTURE

A. *Model-View-Controller (MVC)*

URMIS is based upon the widely used Model-View-Controller (MVC) architecture common in interactive web based applications.

MVC separates the layers; presentation layer (UI: User Interface), business (BLL: Business Logic Layer) and data access (DAL: Data Access Layer). The goal is to have a minimum length between the different layers of the application; and changes made to any layer of the application do not affect other layers.

B. *URMIS architecture*

URMIS is composed of five layers: client layer, mediator layer, service layer, risk management layer and resource layer. The system is based on the following multi agent systems: client agents, mediator agent, service agents, risk agents, incident agent and internet agent. The figure 4 below represents the architecture of URMIS

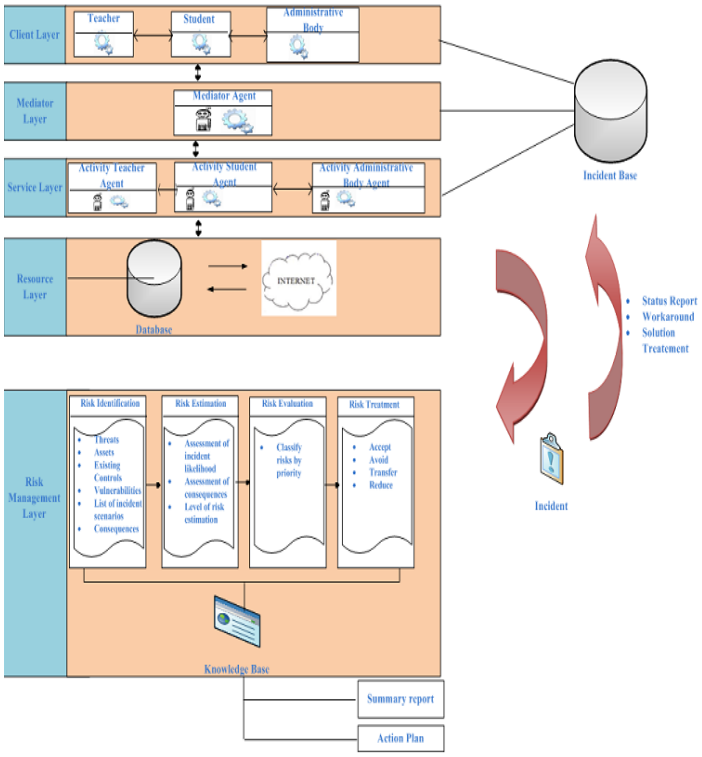


Fig. 3. URMIS architecture

Client agent: They consist of all agents on the client layer, namely agent teacher, agent student, and agent administrative body. They manage the interaction between the users (teachers, students and administrative body) and the system. They allow users to connect to the system by specifying their names, email addresses and id. Every user has a unique id; this field differentiates between the user if it is a teacher, a student or an administrative body.

The id is composed of eight alphanumeric characters; id' teacher starts with the character "t", id's student starts with "s", and id's administrative body starts with "ab". Client agents communicate with the mediator agent by sending users' information of connection.

In case of an incident of connection (password or id forgotten), the user can ask for a solution by sending his request to the risk multi agent system.

Mediator agent: This agent acts like a security checker.

It checks the identities of the users so it can allow them to access to the service layer. It also performs a permission check of the user's access rights and, thereafter, allows him to exploit the service for which he's authorized.

In order to have their needs processed a user requests a service from the mediator agent which then forwards the messages to the destination (service agents) if the access is granted or drops the message and returns a FAILURE message to the sender otherwise. To guarantee a high level of performance several Mediator agents can be triggered to distribute the work among them.

To distinguish between the requesting agents, it is necessary for the mediator agent to link between the user and its category (teacher, student or administrative body). Therefore, the Mediator consults the database which in are stored the users, their corresponding category, and the types of services they can access.

Service agent: These agents communicate with the resource layer to accomplish their tasks.

Risk agents: In the risk management layer we have four agents namely risk identification agent, risk estimation agent, risk evaluation agent and risk treatment agent.

- ✓ ***Risk identification agent:*** It contains in its knowledge base risks that could potentially prevent URMIS to achieve its goals.
  It includes documenting and communicating the users in case of a bad use of URMIS with a list of threats, vulnerabilities and risks that can affect the system.
- ✓ ***Risk estimation agent:*** This agent calculates the likelihood of an incident happening, by applying the risk formula risk=threat * vulnerability * impact.
- ✓ ***Risk evaluation agent:*** It classifies the risk based on the ISO 27005 risk assessment matrix (very low, low, medium, high, very high).
- ✓ ***Risk treatment agent:*** It is in charge of selecting and implementing measures to modify risk. Risk treatment measures can include avoiding, accepting, transferring or reducing risk. The measures (i.e. security measurements) can be selected out of sets of security measurements that are used within the Information Security Management System (ISMS) of the university complying with the standard ISO 27001.

Incident agent: It contains in its knowledge base solutions to similar incidents which occur frequently. This agent stores scenarios of solutions to incidents and make it available for other agents. With this information, other agents are able to take the right decisions in the right moment.

Internet agent: Its role is to store in the knowledge base all the threat and vulnerabilities that it receives from internet.

V. Conclusion and Perspectives

Risk management techniques used before were inappropriate to avoid risks before their occurrence. These approaches were in a reactive perspective. It has therefore become necessary to run into for an integrated approach with a proactive perspective to avoid risks and treat them without compromising the information systems.

In this paper, we describe how information security activities can contribute to the protection of information and infrastructure assets against the risks of loss, bad use or destruction.

In a future work, we will detail the architecture of each agent and the communication between them in URMIS. We will also integrate the processes of the method OCTAVE, in order to quantify risks that can affect URMIS.

References

[1] Y. Rezgui, and A. Marks, "Information security awareness in higher education: An exploratory study," Computers & Security, vol. 27(7-8), pp. 241-253, July 2008.

[2] Defta (Ciobanu) Costinela – Luminita, 2011,Information security in E-learning Platforms, Procedia Social and Behavioral Sciences 15 (2011) 2689–2693

[3] M. Wooldridge. Agents and software engineering. In *AI*IA Notizie* XI(3), 1998 ,pages 31-37.

[4] E. Humphreys ,"Information security management standards: Compliance, governance and risk management".J. Info. Secur. Tech, Rep. 13(4), 247-255,2008.

[5] W. Boehmer, "Appraisal of the effectiveness and efficiency of an Information Security Management System based on ISO 27001". Proc. Second Int. Conf. Emerging Security Information, Sys. & Technologies. pp: 224-231,2008.

[6] A, Kokolakis S, Lambrinoudakis C, Gritzalis S ,"Information Systems Security Management: A Review and a Classification of the ISO Standards". J. Next Generat. Soc. Technol. Leg Issues. 26: 220:35, 2010.

[7] ISO/IEC Guide 73:2002, Risk management – Vocabulary – Guidelines for use in standards.

[8] Consilium-ICT, ITIL et la gouvernance des systèmes d'informations: vers une e administration agile, Toulouse, Juin 2009.

[9] InTech, April 4, 2011. "Multi-Agent Systems - Modeling, Control, Programming, Simulations and Applications", ISBN 978-953-307-174-9

[10] S.Faris,H. Iguer, H.Medromi and S.Sayouti, "New model multi-agent systems based for the security of information system" Proc. IC2INT'13, 2013.

[11] H. Bahtit, B. Regragui.. "Risk Management for ISO27005 Decision Support", International Journal of Innovative Research in Science, Engineering and Technology,2013

[12] S.Faris,H.Medromi and A.Sayouti, "Modélisation d'une plateforme (SIGRCI) à base des systèmes multi-agents & ITIL", JDTIC,2012.

[13] S.Faris,H.Iguer,H.Medromi and A.Sayouti " Conception d'une Plateforme de gestion des risques basée sur les systèmes multi-agents et ISO 27005", JDTIC,2013.

[14] J.Ferber , " Les systèmes multi-agents, vers une intelligence collective InterEditions", 1995

[15] 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" ,intechopen,2011.

[16] A.Sayouti,F.Qrichi Aniba,H.Medromi, "Remote Control Architecture over Internet Based on Multi agent systems". IRECOS, Vol3,N.6,pp.666-671,Novembre 2008.

A Novel Cloud Computing Security Model to Detect and Prevent DoS and DDoS Attack

Masudur Rahman,
Faculty of Business and Services,
Colchester Institute
Colchester,
United Kingdom

Wah Man Cheung
Faculty of Business and Services,
Colchester Institute.
School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

***Abstract*—Cloud computing has been considered as one of the crucial and emerging networking technology, which has been changed the architecture of computing in last few years. Despite the security concerns of protecting data or providing continuous service over cloud, many organisations are considering different types cloud services as potential solution for their business. We are researching on cloud computing security issues and potential cost effective solution for cloud service providers. In our first paper we have revealed number of security risks for cloud computing environment, which has focused on lack of awareness of cloud service providers. In our second paper, we have investigated on technical security issues involved in cloud service environment, where it's been revealed that DoS or DDoS is one of the common and significant dangers for cloud computing environment. In this paper, we have investigated on different techniques that can be used for DoS or DDoS attack, have recommended hardware based watermarking framework technology to protect the organisation from these threats.**

Keywords—Denial of Service attack; Distributed Denial of Service Attack; mechanism of DoS and DDoS attack; framework to prevent DDoS attack, hardware based watermarking

I. INTRODUCTION

Denial of Service (DoS) attack and Distributed Denial of Service Attack (DDoS) are two common types of attacks that do not have a single solution to protect the organisation's IT assets. DoS or DDoS can have severe impact on business and reputation; therefore organisation needs to ensure the security of their IT resources to protect from DDoS attack. By using DoS or DDoS techniques, attacker tries to flood the network or overload the server with traffic so that the legitimate users cannot use the services. Trends to use DoS or DDoS attack have been increased in recent years. These techniques have been used for "cyber warfare" as well. DDoS attack on Visa MasterCard and PayPal by "anonymous" in links to WikiLeaks, DDoS attack on Sony PlayStation, "LulzSec" DDoS attack on CIA and U.K. Serious Organised Crime Agency (SOCA), DDoS attack on WordPress, attack on Hong King Stoke Exchange, CyberBunker DDoS attack are some news, which shows the destructive power of DDoS attack[1].

Successful attacks on these large companies also prove that how vulnerable small organisations are as long as DoS and DDoS are concerned. DDoS attacker may use thousands of different IP addresses to send different types of data packets to the targeted server or network. The process become very complicated for the victim server or network to differentiate between the legitimate traffic and "fake" traffic. Situations become more complicated when the attackers use spoofed IP addresses as source to send the packets, which make it difficult to identify the origins of attacks. The DoS or DDoS attack can cause significant business loss because of less productivity and services, increase downtime; therefore loss in reputation. There are two main reasons that make DoS or DDoS attack very popular among different groups of users. Firstly, there are many tools available to conduct DDoS attack on victim. Most of these tools can be used by attacker without having great deal of technical expertise. Availability of worm maker and ignorance of large number of Internet users make it convenient for attacker to place "bot" into different computers, what can be used for DDoS attack. Secondly, victim organisation will have to spend time and resources to locate attacker, which needs significant involvement of IT security experts[2]. Many organisations are not ready to spend adequate amount of resources to investigate the source of the attack, which encourages the attacker to conduct an attack. Because of the high risk of losing company reputation, number of companies tries not to disclose any security incident in public, which also motivates attacker to use this technique.

In next section we will investigate on different techniques, which can be used in DoS or DDoS attack. After discussing about different DDoS attack, we will propose a framework for cloud computing environment that can use hardware based watermarking technology to detect and prevent DoS or DDoS attack.

II. TECHNIQUES OF DOS / DDOS ATTACK:

Aim of Denial of Service attack is to consume the resources of victim computer's processing power or victim's network bandwidth so that the victim network would not be able to serve legitimate users. This attack generally takes place in very distributed ways, which will make the victim network vulnerable within short period of time. Most of the cases, it is difficult to detect the DoS or DDoS attack early enough to adopt appropriate counter measures to protect the resources because of the distinctive nature and source of this attack. Different types of worms can be used for DDoS attack. This type of attack also can take place in a form of flooding or logic attack. In flooding attack, large amount of "real" but unnecessary data will be send to the victim network or the victim network will receive high volume of request from different sources for specific services. Result of this attack will

consume the bandwidth, processing power or memory of victim network / server; therefore will cause denial of service to legitimate users[3]. Spam emails, data with errors, large volume of data or simple "GET" request for website can cause DDoS attack. Logic DoS attack will be based on exploitation of vulnerabilities within victim system or network. This type of attack needs expertise or intelligent application to identify or to exploit the vulnerabilities of certain networked service[3]. Example of logic attack can be the situation where attacker injects fake routing information to prevent or redirect legitimate traffic from reaching victim's system by exploiting the missing authentication requirements. Sending traffic to the victim system by using fragmented IP datagram can cause system failure, therefore DoS; if the victim computer's operating system or other application software is not securely configured. DDoS bandwidth attack can take place by using TCP SYN flood, ICMP or UDP flood; which will overload the allocated bandwidth of service provider so that legitimate customers will not be able to access their services because of overloaded network. Smurf attacks, Ping of Death attack, TearDrop or Land attack are some common ways to attacking cloud computing environment, which will consume the resources of victim's network and server[4]. Payload in ICMP or DNS reply also can be used for DDoS attack, which will have high probability to pass through the Firewalls. Public Internet Relay Chat (IRC) has been used as tool for DoS attack in recent years by many different attackers.

IP Spoofing is a very popular technique used with DoS attack, where the IP been forged as the traffic coming from victim's network. Alternatively, fake IP can be used as source of the data packet, which does not exist. Upon receiving the data packet, victim system will try to communicate with the forged source system that does not exist. This whole process will consume large amount of resources to cause successful DoS. However, spoofing the IP address of source is not mandatory for DoS or DDoS attack. Attacker may also use number of compromised hosts or chain of proxies to make "trace back" operation difficult to justify the authenticity of source of the packet. Countries with weak or no information security legislation can play significant role to attackers success, if they take the opportunity of this weakness.

R.K. Chang[5] has divided flooding DDoS attack into two main categories names as: direct attack and reflector attack. He has explained direct attack as attackers have spoofed the source IP address and send the traffic and payload directly to the victim computer or network. This type of attack takes place by using ICMP Echo flooding technique, where victim system will have to handle large amount of ICMP Echo request. UDP data flooding is another popular technique used in direct DDoS attack to connect chargen- and echo- ports between two victims. TCP SYN flooding attacks use large number of data packets with forged IP address as source address in packet header so that victim system tries to connect to the source and because of non-exists source address, there will be many "half-open" connection to consume the resources of victim network. Fragmented IP flooding technique also can be used to consume the resources, specifically the memory of the victim server[7].

Unlike this direct attack, reflector attack takes place when the attacker forges the data packet header's IP address. Victim computer's IP will be used as "source address" to send the data payload to a third party by the attacker. On receiving this data packet, third party system will reply to the victim system as that is given as source address in packet header (Fig 1). This type of attack is complicated to response to protect the organisation while this technique can be used to bring more than one system down at the same time. Many different types of networking protocols can be used for reflector DDoS attack, which can include any application layer protocols to request data from web server or DNS server.

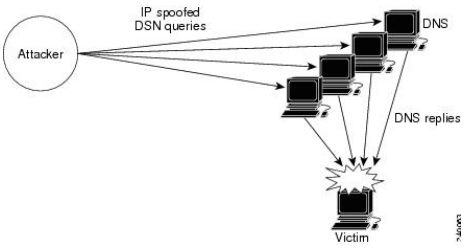


Fig 1: DDoS with Spoofing

This type of attack also can use packet amplification techniques, where third party will reply against a particular request with several packets of data. An example of this type of attack can be, when an ICMP request goes to the broadcast address of a network, therefore each host of that network get the ICMP request. As a result of this request, individual host will reply to the source address to cause DDoS, which has been forged by using victim's IP address. This attack use single ICMP Echo to amplify into many ICMP Echo Reply packets[6].

Flooding attack, both direct and reflector attack can be used against a router to slow down the network or to cause denial of service, which is critical for organisation's network to serve their legitimate users. Using this DDoS attack technique against DNS is a common threats, which cause significant disruption to the organisation's business. Simple tools can be used to perform flooding DDoS attack. This may not requires high volume of resources or bandwidth, which made this type of attack very popular among different attackers. If the flooding attacks take place in a form of distributed attack, attacker must have to have DDoS agents into different systems, which had been compromised before.

In contrast to the nature of flooding attack, logic attack will use different types data packets to exploit the vulnerabilities of victim system. This attack will take place in a form of direct attack, as the attacker will already know the vulnerabilities of the targeted system. Some common methods used in DDoS logic attack include exploitation of syntax or semantic error in victim system. Bugs in the system can be used as vulnerability to attack by attacker (Fig 2).

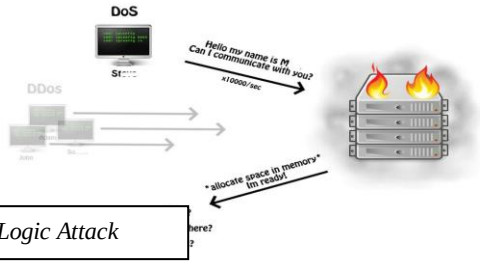


Fig 2: Logic Attack

In next section of this paper, we will be investigating on different mechanisms to detect DoS or DDoS attack.

III. Mechanism To Detect Dos Attack

Different types of techniques and technologies have been used to identify Denial of Service (DoS) or DDoS attach. Incoming traffic can be analysed mainly in two different ways. Manual or automated anomaly detection system can be used to identify DoS attack, where trend of network use can be set against the real network use; therefore any deviation from expected traffic into the network can be identified as sign of DoS or DDoS attack and can be looked into the further details. However, it is a challenge to decide the expected behaviour of the network or users[8]. This type of detection mechanism can detect new or modified attacks including zero days' vulnerabilities. Traffic analysis mechanism to identify DoS or DDoS attack also can work based on signature. Signature based mechanism works in similar principal like antivirus, where the attacks will only be identified if the attack type already given into the detection system. Signature based DoS attack identification mechanism is hugely vulnerable to zero days attack. Attackers also may change or modify the attacks type or tools just to avoid the detection system.

Intrusion Detection System (IDS) are widely been used to identify DoS attack, which normally has three different sections to perform different tasks. Part of IDS will be able to use different sensors to collect data from network or host machine, which will then be analysed to identify abnormal activities and the last function of traditional IDS will be to generate alert for administrator as well as logging the incident for future use. There are two main categories of IDS been used; Network based Intrusion Detection System (NIDS) and Host based Intrusion Detection System (HIDS). NIDS will be able to collect the data from whole network while HIDS will be used to collect the data only from specific host. Having both of these systems in place can provide efficiency in data collection, therefore effective analysis and generating alerts in case of any intrusion attack. Intrusion Detection Systems are necessary part of network now to ensure timely detection of any attacks including DoS or DDoS. Log from IDS is very important to prosecute the attacker. This log also can be used to analyse the attacking method for future attack prevention.

In next section of this paper, we will be investigating on different mechanisms to prevent DoS or DDoS attack.

IV. Mechanism To Prevent Dos And Ddos Attack

There is no single solution against DoS or DDoS attack as the attacker can use many different methods of attack. Organisation will have to adopt different protective mechanisms to have efficient defence against DoS attack. A *defence in depth* approach will help to fight against DoS, where there will be different layers of protection by using different security strategies and technologies. To prevent DoS attack, there needs to be minimum two layers of protective mechanism. First layer will react on *deployment phase* of the attack, when the attacker might try to spread a worm or start the TCP SYN flood to the network. Second layer of defence mechanism will react on time of active attack to prevent the network. In this section, we will be explaining different mechanisms to protect the organisation's IT systems against DoS attack. We also will propose a cost effective potential solution to prevent DDoS attack.

Operating Systems and different applications can raise massive security concerns in terms of being victim of DoS attack, while individual software within a networked system not has been configured efficiently to ensure the optimum security. Unnecessary ports and services can be enabled into a system, which can be used by the attacking tools or attackers. Having updated signature database for antivirus and security patches of OS can contribute significantly in defence mechanism of DoS attack. It is important to ensure that each device within the network does have strong authentication system so that *logic attack* cannot take place by modifying the configuration of router or such other devices. IDS should be used with effective customisation according to the needs of network environment. Both NIDS and HIDS can be used to minimise the false positive and false negative alerts and to detect attacks on early stage. Firewall or similar device should be used for access control to the network. Many resources may needs to be used within the network but not from Internet; therefore access list should be implemented according to the security policy[10]. A support for Quality-of-Service (QoS) features should be configured in router.

If an intrusion has been identified while in *deployment phase of the attack*, there should be an automated mechanism of killing process, restarting application and killing active connections by using TCP RST in case of TCP SYN attack[11]. Propagation of worm typically based on *stack smashing attack*, where attacker managed to access compromised host systems[12]. In terms of propagating worms, a compromised host will establish connection with many different hosts within short period of time; therefore limiting the rate of connections for certain time will be an effective defence mechanism against this type of attack[13]. Vulnerability scanner plays important role for identifying weaknesses within the networked system. Software auditing including checking the vulnerabilities for buffer overflow attack, SQL injection or XSS attack should be conducted regularly to prevent these types of attacks.

In next section of this paper, we will be proposing hardware based watermarking mechanisms to detect and prevent DoS or DDoS attack.

V. Proposed Preventive Mechanism Against Ddos Attack:

One of the complicacy to have effective defence against DDoS is identify the attacked traffic separately than legitimate traffics. Attackers normally use many spoofed IP addresses to attack the system, therefore it become resource consuming to check each of the data packets. Edge routers can be used to mark the source of data packet by using reverse checking mechanism. In case of DDoS attack, large volume of data will be coming from certain hosts. If the source IP has been forged, the type of data will be identical for most of the cases. Using TTL or hop counts, data packets can be grouped as trusted or untrusted. To perform this operation, "hardware based watermarking technology" can be used. This section of the

network will use *traceback mechanism* by using hop count and TTL to test the authenticity of source of data, anomaly of the type of data and classify the source/data as trusted or untrusted. Hardware watermarking also will maintain a table, which can use certain cache timing, so that traffic from certain host is not blocked permanently. After certain time, hardware watermarking will check the incoming data packets from specific host again as that might be a legitimate host machine, which had been compromised because of attackers worm. Updated table of trusted and untrusted hosts should be passed to the next device of the network such as router to send the traffic to the right destination. This device should have defence mechanism to protect itself from DDoS attack especially for TCP SYN attack. Specific TTL should be assigned *traceback operation* to verify the source of information. Having hardware watermarking and filtering technology can work as efficient and cost effective defence mechanism against DDoS attack for any organisation because of less consumption of resources.

Proposed hardware-based watermarking technology to detect and prevent DoS attack will work on following principles:

1) Once packet will reach to the network, source of the packet will be identified.

2) Traceback mechanism should be used to check the authenticity of source address by using Hop Counts and TTL.

3) If the source cannot be verified, packet will be marked as untrusted and will be dropped without sending it to internal network.

4) Each packet coming from same untrusted source will be grouped together based on source authenticity (Figure 3)

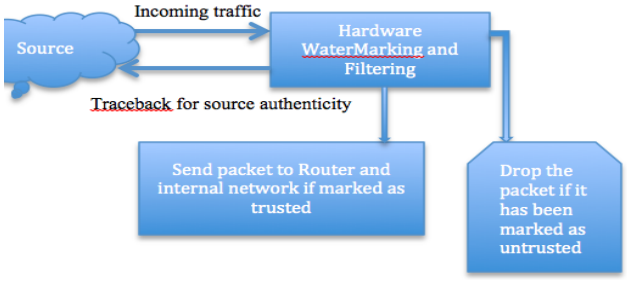


Fig 1: Proposed IDS System

5) If the source is verified, anomaly of the data packets and connection mechanism should be checked against "knowledge based database". Any suspicious data packets sho[illegible]or in depth investigation to reduce the rate of false positive or false negative response.

6) Based on known attack type, packet and source should be marked as untrusted and drop the packet on edge of the network.

7) Only "trusted" packets should be marked and passed to the internal network.

VI. Conclusion and Future Research:

This paper is very beginning of the research to design a cost effective solution for cloud computing environment to prevent DoS attack. In this early stage of this research, we tried to identify the nature and severity of DoS attack and different techniques used by this attack, so that we can design and build the prototype.

Hardware based watermarking and filtering mechanism can provide additional layer of defence against DDoS attack, which also will consume less resources. We will continue this research to present an algorithm, build and test prototype for Hardware based watermarking and filtering method to prevent the network from DDoS attack.

References

[1] http://www.itbusinessedge.com/slideshows/show.aspx?c=92910&slide=7

[2] CERT Coordination Center, Overview of attack trends, Feb 2002 (online)

[3] Dr. Moore, G.M. Voelker and S. Savage, Inferring Internet Deniel of Service Activity.

[4] M. Rahman, W.M. Cheung, Analysis of Cloud Computing Vulnerabilities.

[5] R.K. Chang, Defending against flooding based distributed denial of service attacks: a tutorial, IEEE Commun. Mag.

[6] S. Northcutt & J. Novak, Network Intrusion Detection, Third Edition

[7] B. Guha & B. Mukherjee, Network security via reverse engineering of TCP code: vulnerability analysis and proposed solution

[8] CERT Coordination Center, Overview of attack trends, Oct 1997 (online)

[9] www.cisco.com

[10] M. Handley, V. Paxson and C. Kreibich, Network intrusion detection: evasion, traffic normalisation and end-to-end traffic semantics

[11] V. Paxon, An analysis of using reflector for distributed denial of service attack.

[12] Cisco Systems, Characterising and tracing packet flood using cisco router.

[13] M.M. Willliamson, Throttling Viruses: restricting propagation to defeat malicious mobile code.

Fast Efficient Clustering Algorithm for Balanced Data

Adel A. Sewisy
Computer Science Department
Faculty of Computer and Information, Assiut University
Assiut, Egypt

M. H. Marghny
Computer Science Department
Faculty of Computer and Information, Assiut University
Assiut, Egypt

Rasha M. Abd ElAziz
Computer Science Department
Faculty of Science, Assiut University
Assiut, Egypt

Ahmed I. Taloba
Computer Science Department
Faculty of Computer and Information, Assiut University
Assiut, Egypt

***Abstract*—The Cluster analysis is a major technique for statistical analysis, machine learning, pattern recognition, data mining, image analysis and bioinformatics. K-means algorithm is one of the most important clustering algorithms. However, the k-means algorithm needs a large amount of computational time for handling large data sets. In this paper, we developed more efficient clustering algorithm to overcome this deficiency named Fast Balanced k-means (FBK-means). This algorithm is not only yields the best clustering results as in the k-means algorithm but also requires less computational time. The algorithm is working well in the case of balanced data.**

***Keywords*—Clustering; K-means algorithm; Bee algorithm; GA algorithm; FBK-means algorithm**

I. INTRODUCTION

The problem of clustering is perhaps one of the most widely studied in the data mining and machine learning communities. This problem has been studied by researchers from several disciplines over five decades. Applications of clustering include a wide variety of problem domains such as text, multimedia, social networks, and biological data. Furthermore, the problem may be encountered in a number of different scenarios such as streaming or uncertain data. Clustering is a rather diverse topic, and the underlying algorithms depend greatly on the data domain and problem scenario [1-6].

The goal of clustering is to group a set of objects into clusters so that the objects in the same cluster are highly similar but remarkably dissimilar with objects in other clusters. To tackle this problem, various types of clustering algorithms have been developed in the literature. Among them, the k-means clustering algorithm [7] is one of the most efficient clustering algorithms for large-scale spherical data sets. It has extensive applications in such domains as financial fraud, medical diagnosis, image processing, information retrieval, and bioinformatics [8]. Several clustering algorithms have been developed yet, most of them could not fulfill the requirements of clustering problem which are [9]:

a) High dimensionality: Natural clusters usually do not exist in the full dimensional space, but in the subspace formed by a set of correlated dimensions. Locating clusters in subspaces can be challenging.

b) Scalability: Real world data sets may contain hundreds of thousands of instances. Many clustering algorithms work fine on small data sets, but fail to handle large data sets efficiently.

c) Accuracy: A good clustering solution should have high intra-cluster similarity and low Inter-cluster similarity.

The k-means algorithm and its approaches are known to be fast algorithms for solving such problems. However, they are sensitive to the choice of starting points and can only be applied to small datasets [10].

One common way of avoiding this problem is to use the multi restarting k-means algorithm. However, as the size of the dataset and the number of clusters increase, more and more starting points are needed to get a near global solution to the clustering problem. Consequently the multi restarting k-means algorithm becomes very time consuming and inefficient for solving clustering problems, even in moderately large datasets [11].

In this paper, a new clustering algorithm is proposed for clustering large data sets called FBK-means. The algorithm minimizes an objective function to determine new cluster centers. Compared with the K-means algorithm and other existing modifications, the FBK-means algorithm can obtain a slightly better result but with a lower computational time. The algorithm is working well in the case of balanced data.

The rest of this paper is organized as follows. Section 2 reviews the k-means algorithm and some existing modifications. Section 3 presents a more efficient FBK-means algorithm. Section 4 analyzes the results of the proposed algorithm. Finally, Section 5 concludes the paper with some remarks.

II. BACKGROUND

In this section, we give a brief description of the k-means and some existing modifications.

A. K-means algorithm

K-means algorithm is one of the most popular clustering algorithms and is widely used in a variety of fields. In k-means algorithm, a cluster is represented by the mean value of data points within a cluster and the clustering is done by minimizing the sum of distances between data points and the corresponding cluster centers. Typically, the squared-error (SE) criterion is used, defined as:

$$SE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_j - m_i \right\|. \quad (1)$$

Where SE, is the sum of square-error for all objects in the dataset, k number of clusters, n_i number of objects in each cluster, x_i is the point in space representing a given object, and m_i is the mean of cluster c_i.

The validity of all clusters is defined as [12]:

$$\text{Validity} = \frac{\text{Inter\_Cluster\_Dist}}{\text{Intra\_Cluster\_Dist}}. \quad (2)$$

Where the intra-cluster distance is defined as:

$$Intra\_Cluster\_Dist = \frac{1}{N} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2 \quad (3)$$

Where N is the total number of data points, S_i, i = 1,2,.., k, are the k clusters and μ_i is the centroid or mean point of all the points $x_i \in S_i$. Another measure of cluster performance is the inter-cluster distance, i.e., the distance between clusters. This is calculated by taking the minimum of the distances between each pair of cluster centroids as follows:

$$Inter\_Cluster\_Dist = \min_{j=i+1,\ldots,k} \left(\left| \mu_i - \mu_j \right|^2 \right), \quad i = 1, 2, \ldots, k-1 \quad (4)$$

We take the minimum of the distance between clusters because it is the upper limit of cluster performance and is expected to be maximized. The ratio of intra-cluster distance to inter-cluster distance can serve as an evaluation function for cluster performance. Since we want to maximize the inter-cluster distance and minimize the intra-cluster distance, we want the validity value to be maximized.

The steps of the k-means algorithm are as follows:

Step 1: Choose a seed solution consisting of k centers (not necessarily belonging to A).

Step 2: Allocate data points a ∈ A to its closest center and obtain k-partition of A.

Step 3: Re-compute centers for this new partition and go to Step 2 until no more data points change their clusters.

This algorithm is very sensitive to the choice of a starting point. It converges to a local solution, which can significantly differ from the global solution in many large data sets. The running time of k-means algorithm grows with the increase of the size and dimensionality of the data set. Hence, clustering of large dataset consumes a great time large error.

Many of the methods discussed these problems, but each method has been focusing on a specific problem, the most important of these methods are genetic clustering algorithm (GA) and Bee-clustering algorithm (Bee).

A. Genetic Clustering Algorithm (GA)

Genetic algorithm [13] is a very popular evolutionary algorithm, formatted by simulating the principle of survival of the fittest in natural environment. It mainly include genes coding, fitness calculations, creating the initial population, determine the evolutionary operation etc, which mainly include selection, crossover and mutation.

The steps of the genetic clustering algorithm are as follows:

Step 1: Set the parameters: population size M, the maximum number of iteration T, the number of clusters K, etc.

Step 2: Generate m chromosomes randomly; a chromosome represents a set of initial cluster centers, to form the initial population.

Step 3: According to the initial cluster centers showed by every chromosome, carry out k-means clustering, each chromosome corresponds to once k-means clustering, then calculate chromosome fitness in line with clustering result, and implement the optimal preservation strategy.

Step 4: For the group, to carry out selection, crossover and mutation operator to produce a new generation of group.

Step 5: To determine whether the conditions meet the genetic termination conditions, if meet then withdrawal genetic operation and tum 6, otherwise tum 3.

Step 6: Calculate fitness of the new generation of group; compare the fitness of the best individual in current group with the best individual's fitness so far to find the individual with the highest fitness.

Step 7: Carry out k-means clustering according to the initial cluster center represented by the chromosome with the highest fitness, and then output clustering result.

A. Bee Clustering Algorithm (Bee)

The Bee Algorithm [14] is an optimization algorithm inspired by the natural behaviour of honey bees to find an optimal solution.

The steps of the Bee clustering algorithm are as follows:

Step 1: Generate initial population of solutions randomly (n).

Step 2: Evaluate each solution using the fitness function.

Repeat the following steps until stopping criterion is met:

Step 3: Select the best solutions (m) for neighborhood search

Step 4: Assign more bees to the ones with highest fitness's (e) out of best solutions (m).

Step 5: Select the fittest bee from each patch.

Step 6: Assign the remaining bees for random search and evaluate their fitness.

III. Fast Balanced K-Means Algorithm

In the k-means algorithm, cluster results depends on the random initial centers. Many developed algorithms try to solve this problem but these algorithms rely idea solution it assumes a population of solutions then select the best and try to find another solutions from them. These steps carried out more than once, until get the best solution and this consumes more time with the increase of the size and dimensionality of the datasets.

To solve these problems we proposed a new fast efficient clustering algorithm for clustering large datasets called FBK-means. This algorithm not only obtain a better result but also with a lower computational time.

The idea of this algorithm, first we generate K random centers then assign each point to its closest center to obtain K clusters, second we compute the validity for each cluster $V_i = integer\left(\frac{D_i - AVG}{Rate}\right)$ where $i = 1,2,\dots,k$, D_i is the sum of distance for cluster i, AVG the average distance for all clusters and the Rate=ER*AVG where ER is error rate. When V_i is positive this means there are overlapping between two or more clusters, if V_i is negative this means there are two or more centers in one cluster and if V_i is zero this means the cluster is better.

Depends on V_i, for each iteration we try to improve the validity of all clusters as in equation (2). By moving the center of cluster that has smallest negative V_i to the cluster that has large positive V_i as in Fig. 1, until all V for all clusters equal zero. Finally apply k-means algorithm to the final centers obtained.

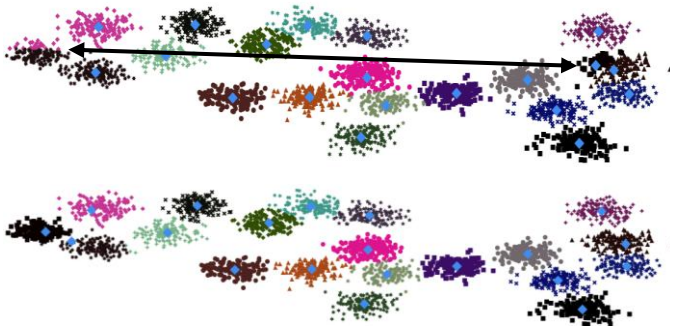


Fig. 1. Moving center from cluster to another for A1 dataset.

The FBK-means algorithm can be summarized as follows:

Step 1: Generate K random centers.

Step 2: For each point a ∈ A Assign its closest center and obtain k clusters of A.

Step 3: Compute the validity of all clusters as in equation (2) and the sum of distance for each cluster D_i, i =1,2,…, k.

Step 4: Compute the average distance for all clusters

$$Avg = \frac{\sum D_i}{k}, i = 1,2,\dots,k$$

Step 5: for each cluster compute:

a- Rate = ER * Avg

b- Validity for each cluster

$$V_i = integer\left(\frac{D_i - AVG}{Rate}\right), i = 1,2,\dots,k$$

Step 6: for each cluster i do one of this:

a) If D_i = 0 (empty cluster) move the center of this cluster to the cluster that has large positive validity.

b) Else if V_i < 0 then moving the center of cluster that has smallest negative validity to the cluster that has large positive validity.

c) Else compute new center to this cluster by computing the average of all of the points of this cluster

Step 7: Repeat Steps 2 to 7 until all V elements equal zero, or the difference between E for each iteration less than the threshold.

Step 8: Apply k-means algorithm to the final centers and compute SE (1) of final results.

Where A is the data, K number of clusters, ER is the error rate, SE is the squared-error, E is the validity of all clusters, D_i, i =1,2,…, k is the sum of distance for each cluster and V_i , i =1,2,…, k is validity of cluster i.

IV. Experimental Results

In order to evaluate the performance of the proposed algorithm, we compare the results with the results of k-means, GA and Bee algorithms on synthetic datasets [15], which are shown in Table I. The implementations have been done in visual studio 2010 on windows7 running on a PC with an Intel core2 duo processor (2.13GHz) with 4GB RAM.

The GA parameters that have been used in the experimental: the population size = 10, selection is roulette, crossover is single point crossover, the probability of crossover = 0.8 and the probability of mutation = 0.001. The Bee parameters that have been used: number of scout bees (n) = 10, number of sites selected for neighborhood searching (m) =5 and number of top-rated (elite) sites among m selected sites (e) =2. For the FBK-means algorithm: the error rate (ER) = 0.2.

TABLE I. Summary of The Datasets

Datasets	Ins. No.	Cluster No.
A1	3000	20
A2	5250	35
A3	7500	50
S1	5000	15
S2	5000	15
S3	5000	15
S4	5000	15
Birch1	100000	100
Birch1	100000	100
Birch1	100000	100

TABLE II. THE AVERAGE SQUARE ERROR AFTER 5 ITERATIONS

Datasets	K-means	GA	Bee	FBK-means
A1	6672474	6314849	6210784	5376830
A2	11808602	11455847	11638661	9251513
A3	16277424	19553389	17054519	13086989
S1	273965201	241313731	247915460	169390458
S2	268798832	270339590	247915460	207120231
S3	274632946	280538768	274874787	241045713
S4	259292277	249340132	240559431	237230705
Birch1	3170853720	3227020055	3115915012	2754436631
Birch2	339313573	306477071	306942942	189361571
Birch3	1830504972	1919705867	1813960089	1621350273

TABLE III. THE AVERAGE TIME IN SECONDS AFTER 5 ITERATIONS

Datasets	K-means	GA	Bee	FBK-means
A1	5	22	67	3
A2	10	69	181	12
A3	64	179	348	19
S1	8	57	64	10
S2	11	68	64	10
S3	7	60	72	16
S4	19	28	67	20
Birch1	418	3094	6845	635
Birch2	297	3070	7005	580
Birch3	778	3599	5980	575

Table II shows the mean square errors after 5 iterations for the k-means algorithm, GA algorithm, Bee algorithm and FBK-means algorithm, it is obviously that the squared error obtained by FBK-means algorithm is better with a lower computational time see Table III.

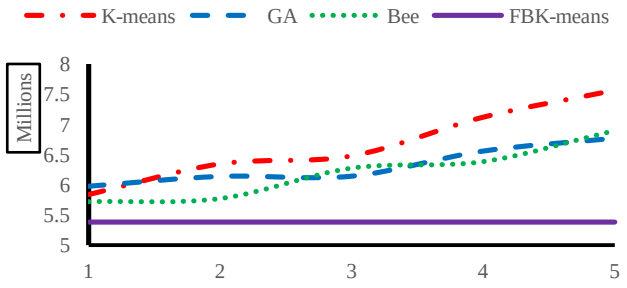


Fig. 2. Square error for A1 after 5 iterations

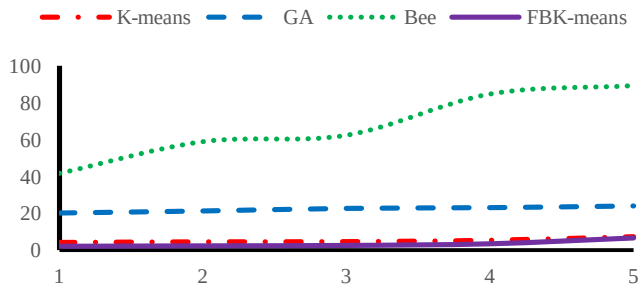


Fig. 3. Time in seconds for A1 after 5 iterations

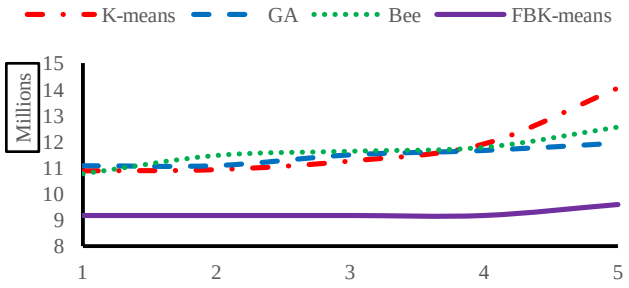


Fig. 4. Square error for A2 after 5 iterations

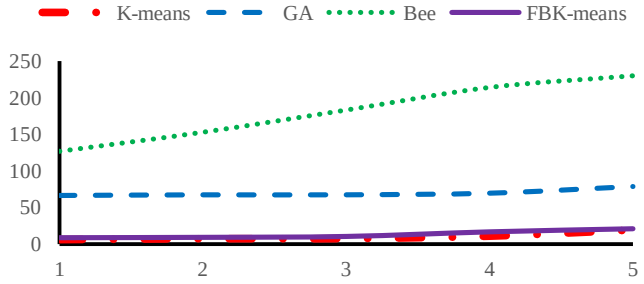


Fig. 5. Time in seconds for A2 after 5 iterations

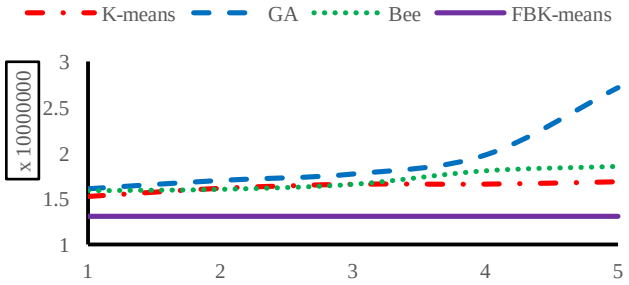


Fig. 6. Square error for A3 after 5 iterations

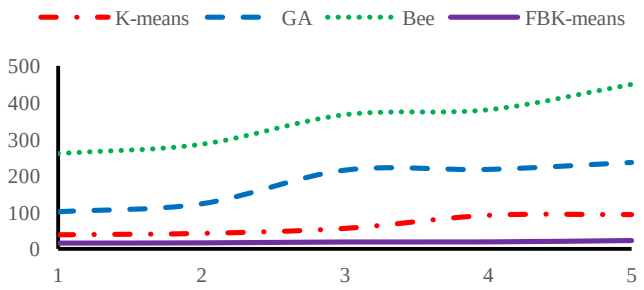


Fig. 7. Time in seconds for A3 after 5 iterations

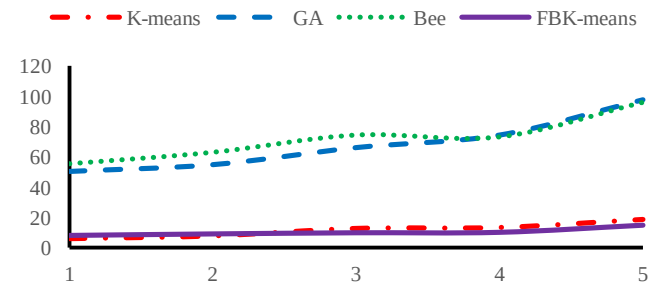


Fig. 11. Time in seconds for S2 after 5 iterations

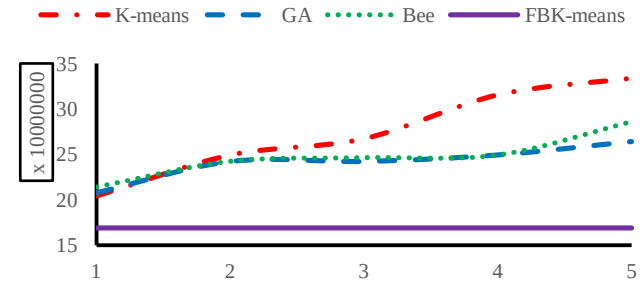


Fig. 8. Square error for S1 after 5 iterations

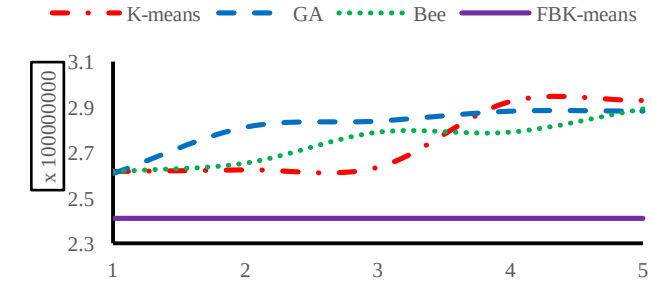


Fig. 12. Square error for S3 after 5 iterations

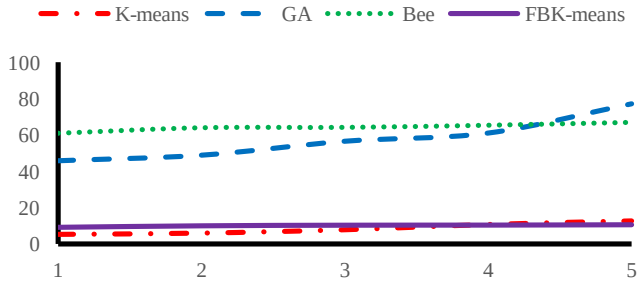


Fig. 9. Time in seconds for S1 after 5 iterations

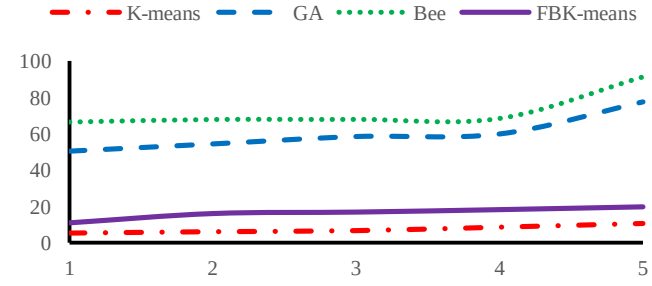


Fig. 13. Time in seconds for S3 after 5 iterations

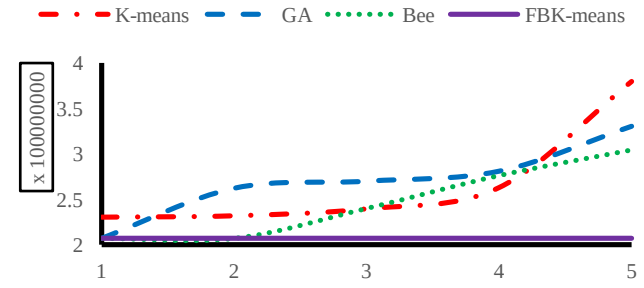


Fig. 10. Square error for S2 after 5 iterations

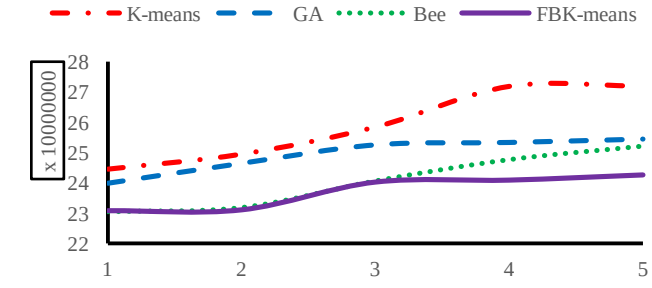


Fig. 14. Square error for S4 after 5 iterations

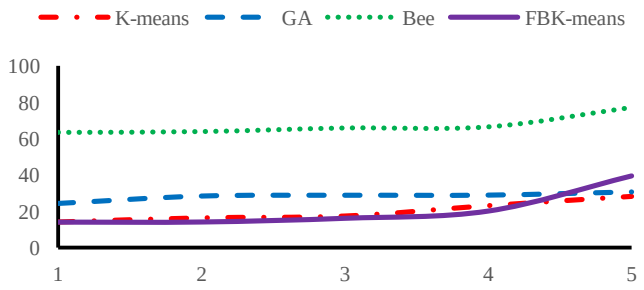


Fig. 15. Time in seconds for S4 after 5 iterations

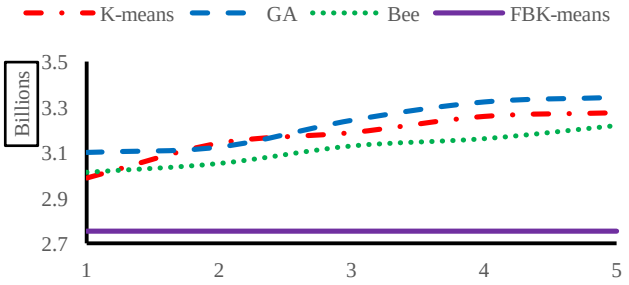


Fig. 16. Square error for Birch1 after 5 iterations

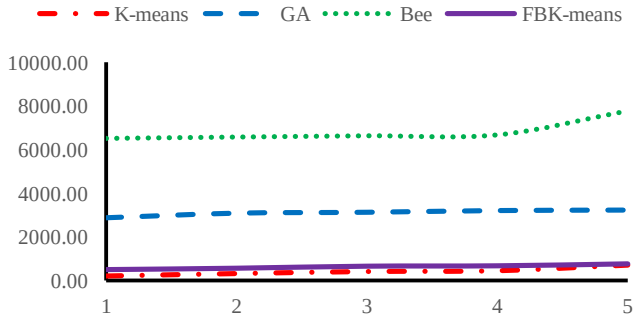


Fig. 17. Time in seconds for Birch1 after 5 iterations

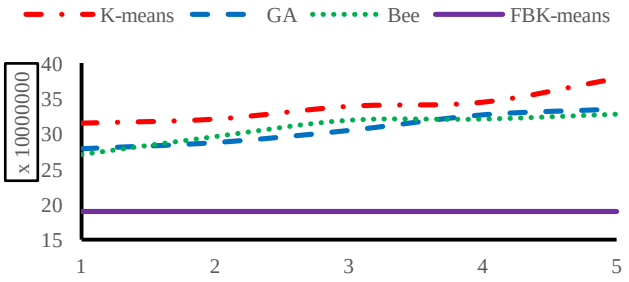


Fig. 18. Square error for Birch2 after 5 iterations

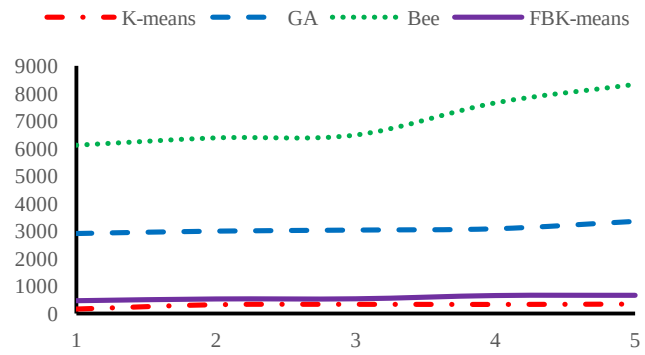


Fig. 19. Time in seconds for Birch2 after 5 iterations

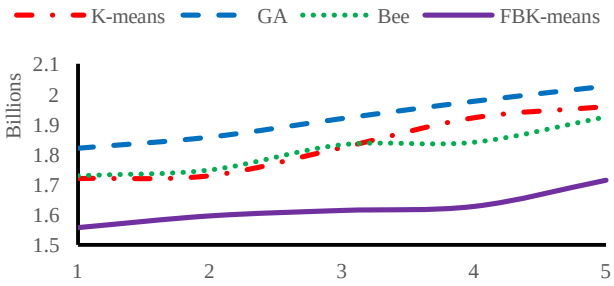


Fig. 20. Square error for Birch3 after 5 iterations

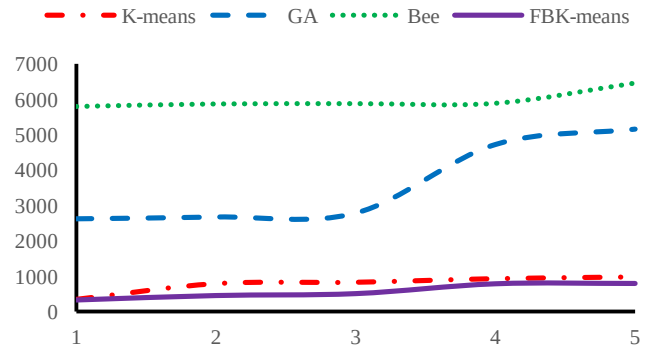


Fig. 21. Time in seconds for Birch3 after 5 iterations.

Fig. 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 shows the mean square errors for ten datasets after 5 iterations, respectively. Fig. 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 shows the execution time for ten datasets after 5 iterations, respectively.

The results show that the FBK-means algorithm outperforms the k-means, GA and Bee algorithms in terms of computing time and square error calculations. When the number of dimensions or clusters k increases, the efficiency of the proposed algorithm becomes more remarkable than the k-means, GA and Bee algorithms.

V. CONCLUSION

To improve the efficiency of the k-means clustering algorithm, a new clustering algorithm is proposed for clustering large datasets called FBK-means. This algorithm minimizes an objective function to determine new cluster centers. Compared with the k-means, GA and Bee algorithms, FBK-means algorithm requires less computing time and fewer distance calculations while retaining the same clustering results. The performance of the proposed algorithm is more remarkable as the number of dimensions or clusters of a dataset increases.

REFERENCES

[1] C.C. Aggarwal and C.K Reddy, "Data clustering: algorithms and applications", Chapman and Hall/CRC Press, 2013.

[2] M.H.Margahny and A.A.Mitwaly, "Fast Algorithm for Mining Association Rules", AIML 05 Conference, Cairo, Egypt, 2005

[3] Marghny, M. H., and A. A. Shakour, "Fast, Simple and Memory Efficient Algorithm for Mining Association Rules", International Review on Computers & Software, 2007.

[4] Margahny, M. H., and A. A. Shakour, "Scalable Algorithm for Mining Association Rules", ICCST, 2006.

[5] M. H. Marghny, Rasha M. Abd El-Aziz and Ahmed I. Taloba, "An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study", Computer Science Department, Egypt, International Journal of Computer Applications, vol. 34, No.6, pp. 0975-8887, 2011.

[6] M. H. Marghny and Ahmed I. Taloba, "Outlier Detection using Improved Genetic K-means", International Journal of Computer Applications, vol. 28, No.11, pp. 33-36, 2011.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

[8] L. Ba, J. Liang, C. Sui and D. Dang, "Fast global k-means clustering based on local geometrical information", Information Sciences, vol. 245, pp.168-180, 2013.

[9] N. M. Abdel-Hamid, M.B. Abdel-Halim and M. W. Fakhr, "Bees algorithm-based document clustering", ICIT The 6th International Conference on Information Technology , 2013.

[10] A. Bagirov, J. Ugon and D. Webb, "Fast modified global k-means algorithm for incremental cluster construction", Pattern Recognition, vol. 44, pp.866–876, 2011.

[11] A. Bagirov, J. Ugon and D. Webb, "A new modified global k-means algorithm for clustering large data sets", ASMDA The XIII International Conference , pp.1-5, 2009.

[12] L. An, H. Xie, M. Chin, Z. Obradovic, D. Smith and V. Megalooikonomou, "Analysis of multiplex gene expression maps obtained by voxelation", Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.23-28, 2008.

[13] W. Min and Y. Siqing, "Improved k-means clustering based on genetic algorithm", Computer Application and System Modeling (ICCASM), vol. 6, pp.636–639, 2010.

[14] D. T. Pham, S. Otri, A. Afify, M. Mahmuddin, and H. Al-Jabbouli, "Data clustering using the bees algorithm", 40th CIRP International Manufacturing Systems Seminar, 2007.

[15] "http://cs.joensuu.fi/sipu/datasets/"

Encrypted With Fuzzy Compliment-Max-Product Matrix in Watermarking

Sharbani Bhattacharya, Associate Professor(IT)
IEC-College of Engineering & Technology, Greater Noida, India

***Abstract*—Watermark is used to protect copyright and to authenticate images. In digital media, today's world images are in electronic form available in the internet. For its protection and authentication invisible watermarking in encrypted form are used. In this paper encryption is done using fuzzy Compliment-Max-Product matrix and then encrypted watermark is embedded in the digital media at desired places using fuzzy rule. The Region of Interest (ROI) is decided with fuzzification. Then, watermark is inserted at the respective positions in the image. Robustness of watermark is judged for ROI. This method of watermarking is done on all image file formats and it is resistant for geometric, noise and compression attack.**

Keywords*—*Watermarking; Fuzzy Compliment-Max-Product Matrix, Fuzzification; Encryption

I. Introduction

Watermark is used to protect copyright and to authenticate images. In digital media, today's world images are in electronic form available in the internet. For its protection and authentication invisible watermark in encrypted form are used. Cryptography is an art of converting a message into cipher text and send to the destination. The authorized person can decipher the text and retrieve the original message. This technology is used from very beginning of the civilization. As the days go by we have new and recent technology coming up. Prior we had texts which are converted into cipher text by using some notion that A should read as B and B should be read as C and so on. Doing this we get a cipher text which is not easily readable unless one knows the conversion method. Public Key and Private Key method is used for cryptography. There are many methods like RSA, DES, Diffie-Hellmann and etc. for cryptography.

In 1976, Martin Hellman, a professor at Stanford University, and Whitfield Diffie, a graduate student, introduced the concept of asymmetric or public key cryptography. In this paper encryption is done using proposed Fuzzy Compliment-Max-Product matrix composition and then encrypted watermark is embedded in the digital media at desired places using fuzzy rule. The Region of Interest (ROI) is decided with fuzzification. Then, watermark is inserted at the respective positions in the image. Robustness of watermark is judged for ROI. Robustness is concerned about tracing or tampering of watermark by attacker. A good watermark should be against filtering process, noise addition, lossy compression, geometry transformation such as rotation, scaling and translation.

The proposed method of watermarking is done on all image file formats and it is resistant for geometric filters, noise and compression attack.

II. Fuzzy Rules and Compliment- Max-Product Matrix

The Fuzzy rules are consisting of rules defined on fuzzy set. Fuzzy set are acquired from Crisp Set using membership function. This process is known as fuzzification. Converting fuzzy set to Crisp set is called defuzzification. In Fuzzy sets the elements are from 0 to 1. Here, we will be using fuzzy Compliment-Max-Product matrix for encryption of the text/file. The encrypted file is then embedded into digital image as watermark. The embedding process also involves fuzzy rules. The encrypted watermark can be extracted from the digital image in unified format. The unified format is then decrypted using algorithm. If the watermarked image is tried to tamper or change, the information can be obtained from the image. The proposed rule is Fuzzy Compliment-Max-Product matrix composition. This rule is consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B(y1,y2) and C(z1,z2).Let us say,

$\mu_{A,B}(x1,y1)=0.2$

$\mu_{A,B}(x1,y2)=0.3$

$\mu_{A,B}(x2,y1)=0.2$

$\mu_{A,B}(x2,y2)=0.4$

$\mu_{B,C}(y1,z1)=0.3$

$\mu_{B,C}(y1,z2)=0.5$

$\mu_{B,C}(y2,z1)=0.2$

$\mu_{B,C}(y2,z2)=0.2$

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}(x1,z1)=1-\max\{[\mu_{A,B}(x1,y1) * \mu_{B,C}(y1,z1)],[\mu_{A,B}(x1,y2) * \mu_{B,C}(y2,z1)]\}=0.94$

$\mu_{A,C}(x1,z2)=1-\max\{[\mu_{A,B}(x1,y1) * \mu_{B,C}(y1,z2)],[\mu_{A,B}(x1,y2) * \mu_{B,C}(y2,z2)]\}=0.90$

$\mu_{A,C}$ (x2,z1)=1- max{[$\mu_{A,B}$ (x2,y1) * $\mu_{B,C}$ (y1,z1)],[$\mu_{A,B}$ (x2,y2) * $\mu_{B,C}$ (y2,z1)]}=0.94

$\mu_{A,C}$ (x2,z2)=1- max{[$\mu_{A,B}$ (x2,y1) * $\mu_{B,C}$ (y1,z2)],[$\mu_{A,B}$ (x2,y2) * $\mu_{B,C}$ (y2,z2)]}=0.90

III. Proposed Encryption Algorithm

The encryption is done using fuzzy set values. The fuzzy rules are then used to decrypt the context. The encryption algorithm has following steps

Step 1: Choose two Fuzzy matrices appropriate for encryption according to the file size.

Step2: Find the Fuzzy Compliment-Max-Product Matrix Composition.

Step3: Generate random number using Fuzzy Matrix.

Step4: Retrieve the encrypted text/files.

There are various ways of encryption. Here, 2X2 fuzzy matrix is used and Compliment-Max-Product of the fuzzy matrix is obtained.

The text/files is encrypted by one of the matrix. File is divided into four parts and a11, a12, a21 and a22. Each part is encrypted using fuzzy matrix values by generating random number using the fuzzy values. The encrypted files are then used for watermarking.

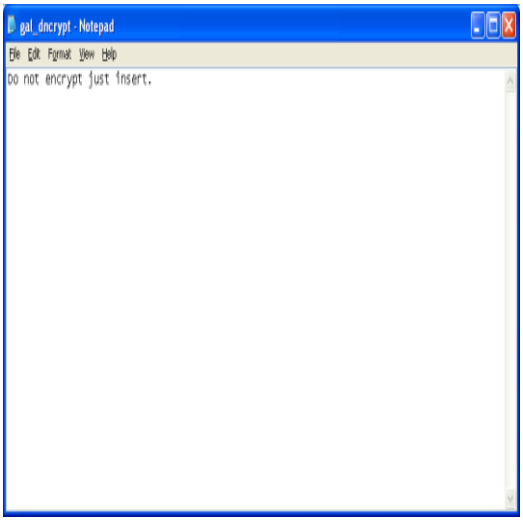


Fig. 1. Text File containing text to be encrypted

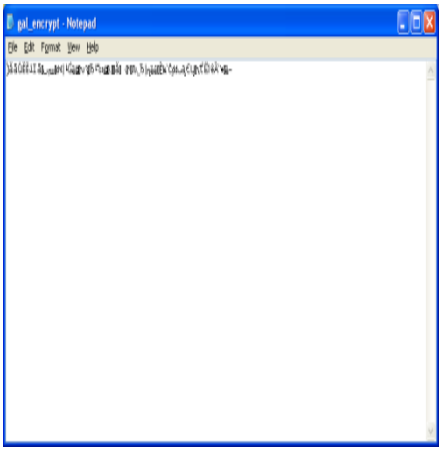


Fig. 2. Encrypted Text to be watermarked

IV. Proposed Decryption Algorithm

Decryption algorithm is used decrypt the encrypted file. The following algorithm is used-

Step1: Choose Fuzzy matrix key for decryption coming from encryption algorithm

Step2: Find the Compliment-Max-Product Matrix and break the file into same four parts with appropriate values of fraction.

Step3: Retrieve the original file

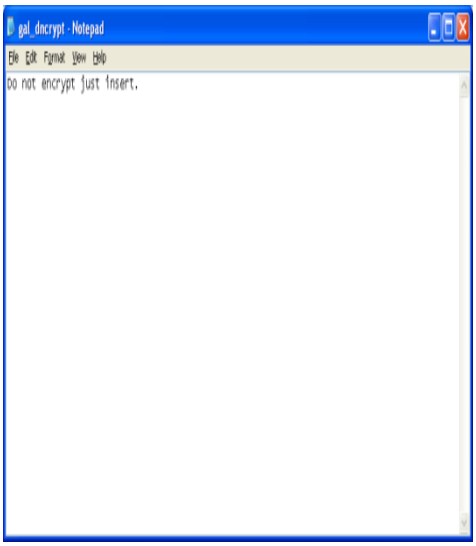


Fig. 3. Decrypted Watermarking File

V. EMBEDDING THE WATERMARK

There are various methods of watermarking. Watermark is visible or invisible. Sometimes invisible watermark is cipher text in order to make watermark more robust and not easily identifiable. When encrypted text or file is embedded in the image the watermark is undetectable. The image is divided into parts one region of interest (ROI) and another is non-region of Interest (i.e. background). Watermarking in the ROI is done where authentication of image is connected.

As we watermark in the ROI then it cannot be easily tampered or deleted. In medical images watermarking is done in non-ROI as to preserve the image accuracy. Here, we will use Fuzzy logic to watermark the image.

The image is having five parts according to ROI

a) *High Priority(HP)*-is the main area
b) *Medium Priority(MP)*- next important
c) *Low Priority(LP)*-least important
d) *Background(BG)*- background of picture
e) Outside (OS)- may be padding zone

All images may not have all five parts. Some may be divided as HP, BG & OS and so on. It depends on image in how many parts it can be divided. OS is padding i.e. the part of file where picture definition is not stored. This is the non-visible region of image.The Fuzzification is converting crisp set to fuzzy set with a membership function. Here, we have Crisp Set A consisting the four image zone. The Fuzzy Set A consists of elements with membership function μ. Here, we want to embed watermark at ROI of image. We take A={HP,MP,LP, BG}. In case of medical image we take B={ BG, OS} as crisp set and then fuzzfication is done.

$A=\{x_1, x_2, x_3, x_4\}$

$\mu_A(x_1)=0.44/$ HP $\mu_A(x_2)=0.42/$ MP

$\mu_A(x_3)=0.14/$ LP $\mu_A(x_4)=0.14/$ BG

The watermark is divided into four fuzzy set elements x_1,x_2 ,x_3 and x_4. It is stored in the regions HP,MP ,LP and BG respectively. Fig 4 shows watermarked image with fuzzified encrypted text file. When we want to retrieve the watermark we need to defuzzify and collect the four parts of watermark then its combined to one file. Then, this file is decrypted using decryption algortihm as given above.Different images have different membership functions for embedding like some images have logic like (0.25/x1, 0.25/x2, 0.25/x3, 0.25/ x4). The x1 is starting of embedding, x2, x3 and x4 are end of ROI of image. In order to make a robust watermarking it should be resistant to noise attack, geometric filter attack and compression attack. Embedding is resistant to all of three above said method.

Steps to be followed for watermarking-

Step1: Deciding the region for embedding using fuzzy membership function.

Step2: Divide encrypted watermark into four files.

Step3 : Convert the Digital Image into Byte Code.

Step4: Convert the Encrypted Watermark files to Byte Code.

Step5: Insert the Byte code of Watermark into Image file using Fuzzy rule.

Step6: Convert the Byte Code to desired Image File Format.

VI. RETRIEVING THE WATERMARK FROM DIGITAL IMAGE

Retrieving the watermark is done by extracting watermark from Digital Image and decrypting the file and obtained the watermark. Steps to be followed are as follows-

Step1: Convert the Watermarked file into Byte Code.

Step2: Extract the Byte Code of Watermark using defuzzification.

Step3: Convert the Byte Code file to text/file.

Step4: Decrypt the text/file using Decryption algorithm.

VII. CONCLUSION

The digital images are watermarked with encrypted files in order to have invisible watermark. The watermark is encrypted and decrypted to see whether image is authentic or it is tried to tamper. The watermark is robust against geometric filter attack, scaling attack, compression attack and noise attack. The drawback of the method is, it uses fraction values for encryption like you encrypt by 0.2. Now for decryption 0.1, 0.2 and 0.3 values can work out. This is loop hole of fractions as values are nearby. So, appropriate programming is required so that decryption cannot be done with other than expected value or key. The proposed method of watermarking depends upon the type of image on which watermark is used. The medical images, images of natural calamity and weather forecasting, company logo, Software logo and etc have different requirements for authenticity and copyright protection. In medical images and weather forecasting ROI is most important so watermark should be out of ROI. In company logo or Software logo tampering or deletion of watermark is an issue so watermark is to be embedded in ROI.

VIII. FUTURE SCOPE OF WORK

The future scope of work is on invisible watermark, fuzzy rules creation and embedding the watermark so that difference in quality of image of original and watermarked image is minimal.

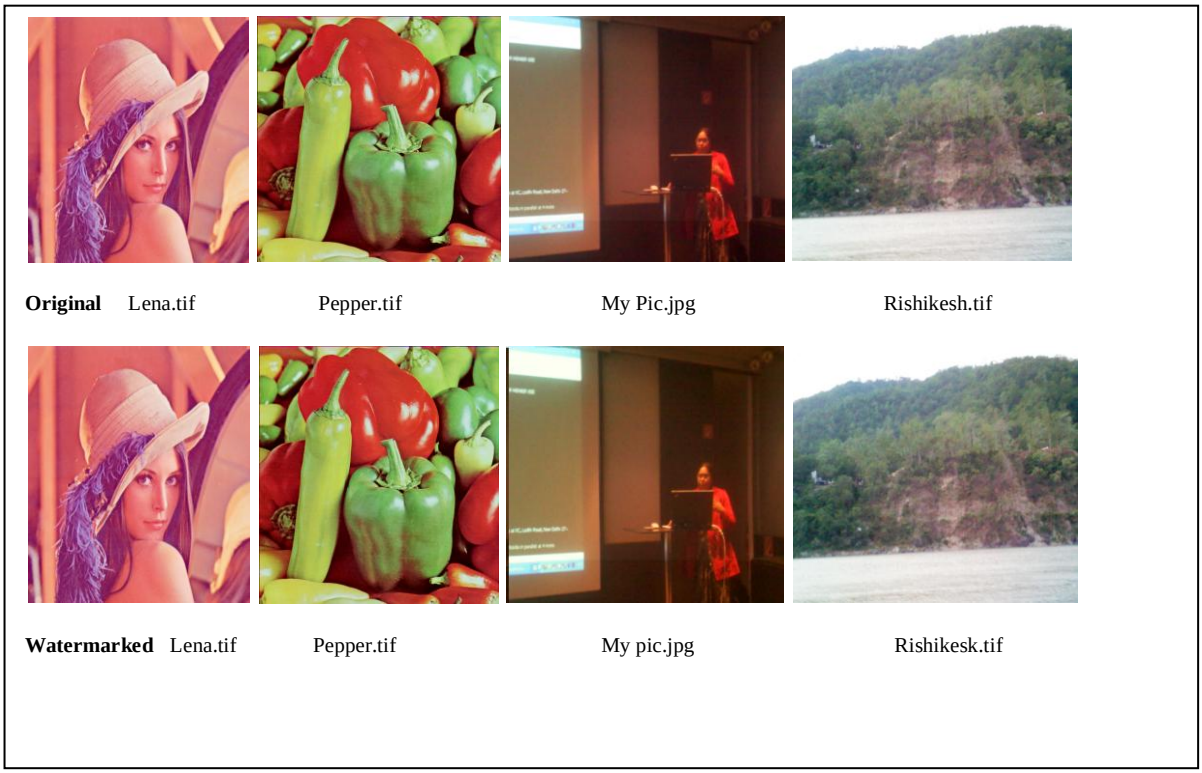


Fig. 4. Original and Watermarked Images

REFERENCES

[1] Anandabrata Pal and Nasir Memon, "Evolution of File Carving",Page 59, IEEE Signal Processing Magazine, March 2009.

[2] Anderson Rocha, Walter Scheirer and Terrance Boult, Siome Goldenstein, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics" ACM Computing Surveys, Vol. 43, No. 4, Article 26, Publication date: October 2011.

[3] Alain Tr ´emeau,Shoji Tominaga and Konstantinos N. Plataniotis, *Review Article* "Color in Image and Video Processing: Most Recent Trends and Future Research Directions", Hindawi Publishing Corporation EURASIP Journal on Image and Video Processing Volume 2008, Article ID 581371, 26 pages ,doi:10.1155/2008/581371.

[4] Alimohammad Latif, " An Adaptive Digital Image Watermarking Scheme using Fuzzy Logic and Tabu Search", Journal of Information Hiding and Multimedia Signal Processing, Volume 4, Number 4, October 2013.

[5] Chuntao WangJiangqun Ni,and Jiwu, "An Informed Watermarking Scheme Using Hidden Markov Model in the Wavelet Domain" Huang,*Page853*, IEEE Transactions On Information Forensics And Security, Vol. 7, No. 3, June 2012.

[6] Chun-Hsiang Huang, Shang-Chih Chuang, Yen-Lin Huang, and Ja-Ling Wu, "Unseen Visible Watermarking: A Novel Methodology for Auxiliary Information Delivery via Visual Contents", *Page193*, IEEE Transactions On Information Forensics And Security, Vol. 4, NO. 2, JUNE 2009.

[7] David J. Coumou, Athimoottil Mathew**,"**A Fuzzy Logic Approach To Digital Image Watermarking" **,** ,Rochester Institute of Technology.

[8] Dong Zheng, Yan Liu, Jiying Zhao, And Abdulmotaleb El Saddik**,** "A Survey of RST Invariant Image Watermarking Algorithms",University of Ottawa, ACM Computing Surveys, Vol. 39, No. 2, Article 5, Publication date: June 2007.

[9] E. C. C. Tsang, Changzhong Wang, Degang Chen, Congxin Wu and Qinghua Hu, "Communication Between Information Systems Using Fuzzy Rough Sets", IEEE Transactions On Fuzzy Systems, Vol. 21, No. 3, June 2013, Page 527.

[10] Elzbieta Zielinska, Wojciech Mazurczyk,Krzysztof Szczypiorski, "Trends in Steganography", Communication of ACM, , Vol 57 ,No. 3, March 2014.

[11] Fawad Ahmed, Farook Sattar, Mohammed Yakoob Siyal, and Dan Yu, "A Secure Watermarking Scheme for Buyer-Seller Identification and Copyright Protection" , Hindawi Publishing Corporation EURASIP Journal on Applied Signal Processing Volume 2006, Article ID 56904, Pages 1–15 ,DOI 10.1155/ASP/2006/56904.

[12] Farid ,"Digital Image Forensic" farid@cs.dartmouth.edu ,www.cs.dartmouth.edu/farid.

[13] Franco Frattolillo,"Watermarking Protocol for Web Context",Page 350, IEEE Transactions On Information Forensics And Security, Vol. 2, No. 3, September 2007.

[14] G. Fahmy,M. F. Fahmy and U. S.Mohammed, *Research Article* "Nonblind and Quasiblind Natural Preserve TransformWatermarking" Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing,Volume 2010, Article ID 452548, 13 pages doi:10.1155/2010/452548.

[15] Hooman Tahayori, Alireza Sadeghian *and* Witold Pedrycz, "Induction of Shadowed Sets Based on the Gradual Grade of Fuzziness", ", IEEE Transactions On Fuzzy System, Vol. 21, No. 5, October 2013,Page 937.

[16] Hua Yuan and Xiao-Ping Zhang,, "Multiscale Fragile Watermarking Based on the Gaussian Mixture Model" *Page 3189*, IEEE Transactions On Image Processing, Vol. 15, No. 10, October 2006.

[17] Jen-Sheng Tsai, Win-Bin Huang, and Yau-Hwang Kuo, "On the Selection of Optimal Feature Region Set for Robust Digital Image Watermarking" , page 735, IEEE Transactions On Image Processing, Vol. 20, No. 3, March 2011.

[18] Mohammad Ali Akhaee,Sayed Mohammad Ebrahim Sahraeian, and Craig Jin, "Blind Image Watermarking Using a Sample Projection Approach" Page 883, IEEE Transactions On Information Forensics And Security, Vol. 6, No. 3, September 2011.

[19] Mriganka Gogoi , H.M.Khalid Raihan Bhuyan ,Koushik Mahanta , Dibya Jyoti Das and Ankita Dutta, " Image and Video based double watermark extraction spread spectrum watermarking in low variance

region", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, July 2013.

[20] Roberto Caldelli, Francesco Filippini, and Rudy Becarelli, *Review Article* "Reversible Watermarking Techniques:An Overview and a Classification" , Hindawi Publishing Corporation EURASIP Journal on Information Security ,Volume 2010, Article ID 134546, 19 pages ,doi:10.1155/2010/134546.

[21] Pankaj U.Lande,S.N. Talbar,G.N. Shinde, "FPGA implementation of image adaptive watermarking using human visual model",ICGST-PDCS,Vol.9,Issue1,Oct. 2009.

[22] Pankaj U.Lande ,Sanjay N. Talbar ,G.N. Shinde, "Robust Image Adaptive Watermarking Using Fuzzy Logic An FPGA Approach", International Journal of Signal Processing, Image Processing and Pattern Recognition ,Vol. 3, No. 4, December, 2010.

[23] Peyman Rahmati, Andy Adler and Thomas Tran , " Watermarking in E-commerce", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, 2013.

[24] Qian Ying , Ren Xue-mei, Huang Ying and Meng Li, " Image Sharpness Metric Based on Algebraic Multi- Grid Method",International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4,April 2014.

[25] Sameh Oueslati and Adnane Cherif and Bassel Solaiman, "Maximizing Strength of Digital Watermarks Using Fuzzy Logic", Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.2, December 2010.

[26] Sameh Oueslati, Adnane Cherif & Basel Solaiman, "A Fuzzy Watermarking Approach Based on the Human Visual System", page 218, International Journal Of Image Processing (IJIP), Volume (4): Issue (3) 2010.

[27] Sharbani Bhattacharya, "Watermarking Digital Image Using Fuzzy Matrix Rules", presented in National Conference Smarter Approaches in Computing , Technology & Applications SACTA 2014 at ITS ,Mohan Nagar, Ghaziabad, published in Conference Proceedings Page No 343, 19th April 2014.

[28] S. P. Mohanty, O. B. Adamo, and E. Kougianos, "VLSI Architecture of an Invisible Watermarking Unit for a Biometric-Based Security System in a Digital Camera", in Proceedings of the 25th IEEE International Conference on Consumer Electronics (ICCE), 2007, pp. 485-486.

[29] T.Sridevi and S Sameena Fatima,"Digital Image Watermarking using Fuzzy Logic approach based on DWT and SVD", International Journal of Computer Applications ,Vol 74– No.13, July 2013.

[30] Teresa Garcia-Valverde, Alberto Garcia-Sola, Hani Hagras, James A. Dooley,Victor Callaghan and Juan A. Botia, "A Fuzzy Logic-Based System for Indoor Localization Using WiFi in Ambient Intelligent Environments", IEEE Transactions On Fuzzy System, Vol. 21, No. 4, August 2013, Page 702.

Watermarking Digital Image Using Fuzzy Matrix Compositions and Rough Set

Sharbani Bhattacharya
Associate Professor (IT),
IEC College of Engineering & Technology, Greater Noida , India

***Abstract*—Watermarking is done in digital images for authentication and to restrict its unauthorized usages. Watermarking is sometimes invisible and can be extracted only by authenticated party. Encrypt a text or information by public – private key from two fuzzy matrix and embed it in image as watermark. In this paper we proposed two fuzzy compositions Product-Mod-Minus, and Compliment-Product-Minus. Embedded watermark using Fuzzy Rough set created from fuzzy matrix compositions.**

Keywords*—*Fuzzy Product-Mod-Minus Matrix; Fuzzy Compliment-Product-Minus Matrix; Fuzzy Rough set; Watermarking; Encrypting

I. INTRODUCTION

Watermarking in digital image is for authentication and restricting it for unauthorized usages. Videos and other digital contents are used often by unauthorized users. Watermarking and fingerprinting are used to find the point of leakage or user who allowed unauthorized use of the digital content. The watermarking can be visible or invisible. Visible watermarking is used for authentication whereas invisible watermarking is used for restricting unauthorized usages. The Robustness of watermark depends upon its tolerance towards its tamper or delete. It should be identified and extracted to receive information by authorized party. The amount of embedding of information in a digital content without getting identified is its Capacity.

II. ENCRYPTION METHODS

Cryptography is an art of converting a message into cipher text and send to the destination. The authorized person can decipher the text and retrieve the original message. This technology is used from very beginning of the civilization. As the days go by we have new and recent technology coming up so the cryptographic methods are also changing. Prior we had texts which are converted into cipher text by using some notion that A should read as B and B should be read as C and so on. Doing this we get a cipher text which is not easily readable unless one knows the conversion method. Public Key and Private Key method is used for cryptography. There are many methods like RSA, DES, Diffie-Hellmann and etc. for cryptography. In 1976, Martin Hellman, a professor at Stanford University, and Whitfield Diffie, a graduate student, introduced the concept of asymmetric or public key cryptography. We will use here public key and private key as two fuzzy matrices. One matrix is given by user is public key and second fuzzy matrix is randomly chosen from the database to give resultant private key fuzzy matrix.

III. FUZZY MATRIX COMPOSITIONS

The Fuzzy rules are consisting of rules defined on fuzzy set. Fuzzy set are acquired from Crisp Set using membership function. This process is known as fuzzification. Converting fuzzy set to Crisp set is called defuzzification. Fuzzy set has members which can take values 0 to 1. Thus, Fuzzy set A values like A= {0.1/x1, 0.3/x2, 0.4/x3}. This means 0.1 is membership value of x1 in set A, 0.3 membership value for x2 and 0.4 membership value for x3 in set A. Here, we will be using fuzzy matrix for encryption of the text/file which is to be used for watermarking. The encrypted file is then embedded into digital image using Fuzzy Rough sets. Fuzzy Rough set is P={ inf(A),Sup(A)} where A is the Fuzzy set. Inf(A) is 0.1 i.e. lower bound of set Fuzzy A and Sup(A) is 0.4 i.e. upper bound of set Fuzzy A. The encrypted watermark can be extracted from the digital image in unified format. The unified format is then decrypted using algorithm. The paper proposes two fuzzy matrix composition Fuzzy Product-Mod-Minus composition and Fuzzy Compliment-Product-Minus composition. Embedding will be done creating Fuzzy Rough set from these two new compositions, published Fuzzy Max-Mod-Minus composition and Fuzzy Compliment-Sum-Minus composition [29] and the Fuzzy Max-Min composition.

A. Fuzzy Product-Mod-Minus Composition

The Fuzzy Product-Mod-Minus composition is proposed rule consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B(y1,y2) and C(z1,z2).Let us say,

$\mu_{A,B}(x1,y1)=0.2$

$\mu_{A,B}(x1,y2)=0.3$

$\mu_{A,B}(x2,y1)=0.2$

$\mu_{A,B}(x2,y2)=0.4$

$\mu_{B,C}(y1,z1)=0.3$

$\mu_{B,C}(y1,z2)=0.5$

$\mu_{B,C}(y2,z1)=0.2$

$\mu_{B,C}(y2,z2)=0.2$

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}(x1,z1)= \{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z1)|* | \mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z1) |\}=0.01$

$\mu_{A,C}(x1,z2)= \{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z2) |*|\mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z2)|\}=0.03$

$\mu_{A,C}(x2,z1)= \{|\mu_{A,B}(x2,y1) - \mu_{B,C}(y1,z1)|* |\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z1)|\}=0.02$

$\mu_{A,C}(x2,z2)=\{|\mu_{A,B}(x2,y1) -\mu_{B,C}(y1,z2)|* |\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z2)|\}=0.06$

B. Fuzzy Compliment-Product-Minus Composition

The Fuzzy Compliment-Product-Minus composition is proposed rule consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B(y1,y2) and C(z1,z2).Let us say,

$\mu_{A,B}(x1,y1)=0.2$

$\mu_{A,B}(x1,y2)=0.3$

$\mu_{A,B}(x2,y1)=0.2$

$\mu_{A,B}(x2,y2)=0.4$

$\mu_{B,C}(y1,z1)=0.3$

$\mu_{B,C}(y1,z2)=0.5$

$\mu_{B,C}(y2,z1)=0.2$

$\mu_{B,C}(y2,z2)=0.2$

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}(x1,z1)= |1-\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z1)|*| \mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z1) |\}|=0.99$

$\mu_{A,C}(x1,z2)= |1-\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z2) |*|\mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z2)|\}|=0.97$

$\mu_{A,C}(x2,z1)= |1-\{|\mu_{A,B}(x2,y1) - \mu_{B,C}(y1,z1)| *|\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z1)|\}|=0.98$

$\mu_{A,C}(x2,z2)= |1-\{|\mu_{A,B}(x2,y1) -\mu_{B,C}(y1,z2)|* |\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z2)|\}|=0.94$

C. Fuzzy Max-Mod-Minus Composition

The Fuzzy Max-Mod-Minus composition rule [29] is consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B(y1,y2) and C(z1,z2).Let us say,

$\mu_{A,B}(x1,y1)=0.2$

$\mu_{A,B}(x1,y2)=0.3$

$\mu_{A,B}(x2,y1)=0.2$

$\mu_{A,B}(x2,y2)=0.4$

$\mu_{B,C}(y1,z1)=0.3$

$\mu_{B,C}(y1,z2)=0.5$

$\mu_{B,C}(y2,z1)=0.2$

$\mu_{B,C}(y2,z2)=0.2$

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}(x1,z1)= \max\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z1)|, | \mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z1) |\}=0.1$

$\mu_{A,C}(x1,z2)= \max\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z2) |,|\mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z2)|\}=0.3$

$\mu_{A,C}(x2,z1)= \max\{|\mu_{A,B}(x2,y1) , \mu_{B,C}(y1,z1)|, |\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z1)|\}=0.2$

$\mu_{A,C}(x2,z2)= \max\{|\mu_{A,B}(x2,y1) ,\mu_{B,C}(y1,z2)|, |\mu_{A,B}(x2,y2) - \mu_{B,C}(y2,z2)|\}=0.3$

D. Fuzzy Compliment-Sum-Minus Composition

The Fuzzy Compliment-Sum-Minus composition [29] is consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B (y1, y2) and C(z1,z2).Let us say,

$\mu_{A,B}(x1,y1)=0.2$

$\mu_{A,B}(x1,y2)=0.3$

$\mu_{A,B}(x2,y1)=0.2$

$\mu_{A,B}(x2,y2)=0.4$

$\mu_{B,C}(y1,z1)=0.3$

$\mu_{B,C}(y1,z2)=0.5$

$\mu_{B,C}(y2,z1)=0.2$

$\mu_{B,C}(y2,z2)=0.2$

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}(x1,z1)= |1-\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z1)|+| \mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z1) |\}|=0.8$

$\mu_{A,C}(x1,z2)= |1-\{|\mu_{A,B}(x1,y1) -\mu_{B,C}(y1,z2) |+|\mu_{A,B}(x1,y2) - \mu_{B,C}(y2,z2)|\}|=0.6$

$\mu_{A,C}$ (x2,z1)= |1-{|$\mu_{A,B}$ (x2,y1) - $\mu_{B,C}$ (y1,z1)| +|$\mu_{A,B}$ (x2,y2) - $\mu_{B,C}$ (y2,z1)|}|=0.7

$\mu_{A,C}$ (x2,z2)= |1-{|$\mu_{A,B}$ (x2,y1) -$\mu_{B,C}$ (y1,z2)|+ |$\mu_{A,B}$ (x2,y2) - $\mu_{B,C}$ (y2,z2)|}|=0.5

E. Fuzzy Max-Min Composition

The Fuzzy Max-Min composition is consisting of following method. Let A, B and C are fuzzy set with A(x1, x2), B(y1,y2) and C(z1,z2).

Let us say,

$\mu_{A,B}$(x1,y1)=0.2

$\mu_{A,B}$ (x1,y2)=0.3

$\mu_{A,B}$ (x2,y1)=0.2

$\mu_{A,B}$ (x2,y2)=0.4

$\mu_{B,C}$ (y1,z1)=0.3

$\mu_{B,C}$ (y1,z2)=0.5

$\mu_{B,C}$ (y2,z1)=0.2

$\mu_{B,C}$ (y2,z2)=0.2

The matrix of $\mu_{A,C}$ is

$\mu_{A,C}$ (x1,z1)= max{min[$\mu_{A,B}$(x1,y1) ,$\mu_{B,C}$ (y1,z1)], min[$\mu_{A,B}$ (x1,y2) , $\mu_{B,C}$ (y2,z1)]}=0.2

$\mu_{A,C}$ (x1,z2)= max{min[$\mu_{A,B}$(x1,y1) ,$\mu_{B,C}$ (y1,z2)], min[$\mu_{A,B}$ (x1,y2) , $\mu_{B,C}$ (y2,z2)]}=0.2

$\mu_{A,C}$ (x2,z1)= max{min[$\mu_{A,B}$ (x2,y1) , $\mu_{B,C}$ (y1,z1)], min[$\mu_{A,B}$ (x2,y2) , $\mu_{B,C}$ (y2,z1)]}=0.2

$\mu_{A,C}$ (x2,z2)= max{min[$\mu_{A,B}$ (x2,y1) ,$\mu_{B,C}$ (y1,z2)], min[$\mu_{A,B}$ (x2,y2) , $\mu_{B,C}$ (y2,z2)]}=0.2

IV. Proposed Public Key- Private Key Encryption Algorithm

The encryption is done using fuzzy set values. The fuzzy rules are then used to decrypt the context. The public key is given by user is a fuzzy matrix. There is also a database consisting of fuzzy matrices. The public key given by user and fuzzy matrix from database is chosen randomly and Fuzzy Compliment-Product-Minus composition is used get resultant fuzzy matrix. The encryption is done using this fuzzy matrix. The encryption algorithm has following steps

Step 1: User chooses one Fuzzy matrix appropriate for encryption. It is public key.

Step2: Select one fuzzy matrix from database.

Step3: Find the Fuzzy Compliment-Product-Minus matrix.

Step3: Generate random number using Fuzzy Compliment-Product-Minus matrix.

Step4: Retrieve the encrypted text/files.

There are various ways of encryption. Here, 2X2 fuzzy matrices are used to obtain Compliment-Product-Minus of the fuzzy matrix. The text/files is encrypted by Compliment-Product-Minus of the fuzzy matrix.

V. Proposed Decryption Algorithm

Decryption algorithm is used decrypt the cipher text file. The following algorithm is used-

Step1: Collect the encrypted four parts from four different embedded region of image and combine to for one file.

Step2: Use private key fuzzy matrix key for decryption.

Step3: Retrieve the original file.

VI. Embedding The Watermark

The watermark embedding process is done by inserting the encrypted watermark at appropriate place. Encrypted file is divided into one, two or more parts say b1, b2 and b3.

The three encrypted files are embedded in digital image as watermark using appropriate fuzzy rule. There are many fuzzy matrix compositions like Max-Min, Max-Max, Min-Max, Max-Product, Min-Product composition and etc. The two fuzzy matrices obtained are first used for encrypting watermark. Now for embedding the various compositions of fuzzy matrices are obtained. Using five fuzzy matrix compositions of Section II we will create Fuzzy Rough set P11, P12, P21 and P22. The encrypted three parts of file are inserted at three places of digital image using the most suitable Fuzzy Rough set say P21 and P22. Fig. 4, 5 and 6 shows watermarked images using the fuzzy matrix composition and Fuzzy Rough set.

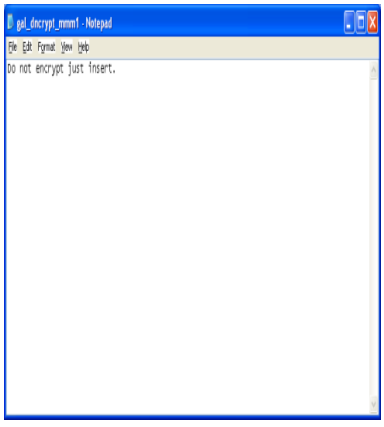


Fig. 1. Text file

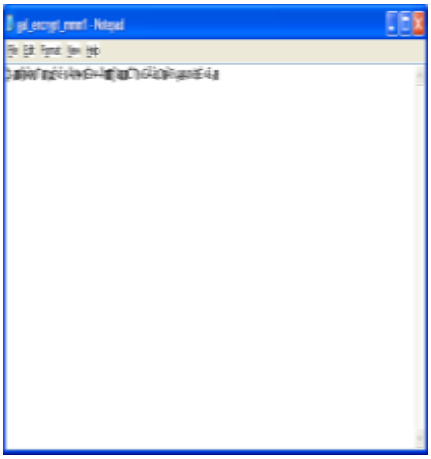


Fig. 2. Encrypted Text

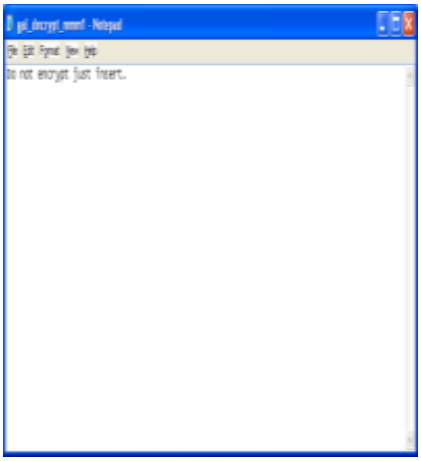


Fig. 3. Decryption Text

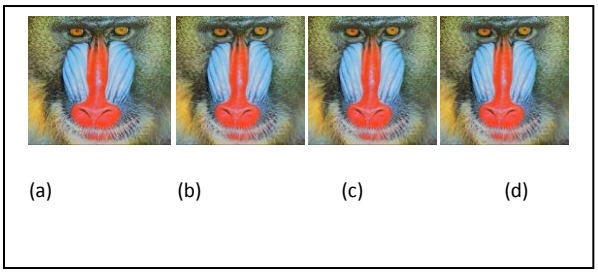


Fig. 4. (a) Original baboon.jpg 4(b) watermark at P22 {sup} 4(c) watermark at P22{inf, sup}4(d) watermark at P22{inf,sup}and P21{sup}

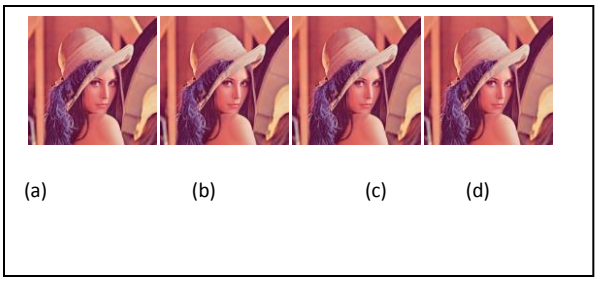


Fig. 5. (a) Original lena.jpg 5(b) watermark at P22 {sup} 5(c) watermark at P22{inf, sup}5(d) watermark at P22{inf, sup}and P21{sup}

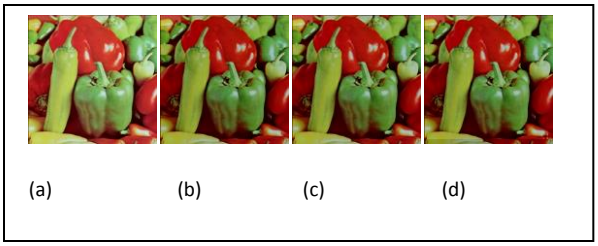


Fig. 6. (a) Original Peppers.jpg 6(b) watermark at P22 {sup} 6(c) watermark at P22{inf, sup}6(d) watermark at P22{sup}andP21{inf, sup}

There are various compositions of fuzzy matrices as said. The most appropriate composition is chosen for

Presently we have used hit and trial method to embed.

Any image can have invisible watermark with fuzzy matrix compositions [29]. Next section explains the proposed method of creating Fuzzy Rough set to embed the watermark.

VII. Proposed Fuzzy Rough Set for Embedding

The upper bound and lower bound concept Rough Set (Pawlak 1991) are used in fuzzy logic gives Fuzzy Rough Set. Fuzzy Rough set are two infimum and supremum chosen from fuzzy values. Let us say Fuzzy set A={0.21/x1 ,0.33/x2,0.3/x3 , 0.4/x4}. The set has infimum 0.2(i.e. lower bound) and supremum is 0.4 (i.e. upper bound) i.e. no element in the set is less than 0.2 and greater than 0.4 respectively. The Fuzzy Rough set is P= {0.2, 0.4}. Two fuzzy matrices are obtained one from database and one by user. From these two fuzzy matrices using five fuzzy matrix compositions of Section II we obtain five Fuzzy matrices i.e. Fuzzy Max-Mod-Minus matrix, Fuzzy Max-Product-Minus matrix, Fuzzy Max-Min matrix, Fuzzy Compliment-Sum-Minus matrix and Fuzzy Compliment-Product-Minus matrix. We create Fuzzy Rough set from five fuzzy matrices. Let us denote

$f_1 = \mu_{A,C}(x1,z1)$,

$f_2 = \mu_{A,C}(x1,z2)$,

$f_3 = \mu_{A,C}(x2,z1)$,

$f_4 = \mu_{A,C}(x2,z2)$.

We get f_1, f_2, f_3 and f_4 fuzzy matrix elements from all five above said fuzzy matrix compositions.

We denote f_{1MPM}, f_{1MM}, f_{1CSM}, f_{1CPM} and f_{1MMM} for $\mu_{A,C}(x1,z1)$ of Fuzzy Max-Product-Minus matrix , Fuzzy Max-Min matrix ,Fuzzy Compliment-Sum-Minus matrix ,Fuzzy Compliment-Product-Minus matrix and Fuzzy Max-Mod-Minus matrix respectively. Similarly, we have f_{2MPM} , f_{2MM} , f_{2CSM} ,f_{2CPM} and f_{2MMM} for $\mu_{A,C}(x1,z2)$ of Fuzzy Max-Product-Minus matrix , Fuzzy Max-Min matrix ,Fuzzy Compliment-Sum-Minus matrix ,Fuzzy Compliment-Product-Minus matrix and Fuzzy Max-Mod-Minus matrix respectively. We have f_{3MPM} , f_{3MM} , f_{3CSM} ,f_{3CPM} and f_{3MMM} for $\mu_{A,C}(x2,z1)$ of Fuzzy Max-Product-Minus matrix , Fuzzy Max-Min matrix ,Fuzzy Compliment-Sum-Minus matrix ,Fuzzy Compliment-Product-Minus matrix and Fuzzy Max-

Mod-Minus matrix respectively. We have f_{4MPM} , f_{4MM} , f_{4CSM} ,f_{4CPM} and f_{4MMM} for $\mu_{A,C}$ (x2,z2) of Fuzzy Max-Product-Minus matrix , Fuzzy Max-Min matrix ,Fuzzy Compliment-Sum-Minus matrix ,Fuzzy Compliment-Product-Minus matrix and Fuzzy Max-Mod-Minus matrix respectively.

We get fuzzy set G11={$0.2/f_{1MPM}$, $0.8/f_{1MM}$,$0.99 /f_{1CSM}$,$0.01/f_{1CPM}$, $0.1/f_{1MMM}$ } , G12={$0.2/f_{2MPM}$,$0.6/f_{2MM}$,$0.98/f_{2CSM}$, $0.03/f_{2CPM}$, $0.3/f_{2MMM}$ }, G21={$0.2/f_{3MPM}$, $0.7/f_{3MM}$,$0.97/f_{3CSM}$, $0.02/f_{3CPM}$,$0.2/f_{3MMM}$ } and G22={ $0.2/f_{4MPM}$,$0.5/f_{4MM}$,$0.94/f_{4CSM}$,$0.06/f_{4CPM}$,$0.3/f_{4MMM}$ } from the all five fuzzy composition matrices. Now, Fuzzy Rough set of each G11, G12, G21 and G22 are P11, P12, P21 and P22 i.e. pair of infimum (inf) and supremum (sup). P11={0.01,0.99} from G11 , P12={0.03,0.98} from G12, P21={ 0.02,0.97} from G21 and P22={0.06,0.94} from G22. We will use these values for embedding watermark at may be at eight points or any four points or three points or so in the image according to our requirement of robustness and invisibility.

Embedding algorithm proposed using Fuzzy Matrices and Rough set

Step1: Get encrypted file and divide it into four or more parts.

Step2: Obtain the two matrices one from user and another from database.

Step3: Obtain the fuzzy matrices using the Fuzzy Max-Mod-Minus composition, Fuzzy Max-Product-Minus composition, Fuzzy Max-Min composition, Fuzzy Compliment-Sum-Minus composition and Fuzzy Compliment-Product-Minus composition.

Step4: Obtain P11, P12, P21 and P22 Fuzzy Rough set from Fuzzy Max-Mod-Minus matrix, Fuzzy Max-Product-Minus matrix, Fuzzy Max-Min matrix, Fuzzy Compliment-Sum-Minus matrix and Fuzzy Compliment-Product-Minus matrix.

Step5 : Embed the watermark at points in the image using P11,P12, P21 and P22 Fuzzy Rough set.

We can break into two or more parts the encrypted file and embedded watermark at P11 {inf(G11) , sup(G11)}, P12 {inf(G12) , sup(G12)} and so on. We get eight points where we can embed watermark from four Fuzzy Rough set. We can also use only one Fuzzy Rough set for embedding at one point say at sup or at inf. We have embedded watermark in three images baboon.jpg, lena.jpg and peppers.jpg using Fuzzy Rough set in Figure 4, 5 and 6. We then find the Peak Signal to Noise Ratio of original image and watermarked image Table I. We obtained the results that in most of the cases PSNR are either or less than 30 Decibel. Only in baboon.jpg when we have inserted at three points have PSNR 42 Decibel. Thus, results show that there is very less deterioration of quality of image (PSNR 30-50) and we also have robustness of watermark. The PSNR 0 means no difference in quality between original and watermarked image.

TABLE I. PSNR OF ORIGINAL AND WATERMARKED IMAGE.

Original Image	**Fuzzy rough set used for insertion of watermark in image**	**PSNR(Db)**
Baboon.jpg Fig 4(a)	1 point of insertion , P22 {sup}, Fig 4(b)	0
Baboon.jpg, Fig 4(a)	2 points of insertion , P22{inf, sup}, Fig 4(c)	0
Baboon.jpg, Fig 4(a)	3 points of insertion , P22{inf, sup} & P21{sup}, Fig 4(d)	42.1102
Lena.jpg, Fig 5(a)	1 point of insertion , P22 {sup}, Fig 5(b)	0
Lena.jpg, Fig 5(a)	2 points of insertion , P22{inf, sup}, Fig 5(c)	0
Lena.jpg, Fig 5(a)	3 points of insertion , P22{inf, sup}, & P21{sup}, Fig 5(d)	0
Peppers.jpg, Fig 6(a)	1 point of insertion , P21{inf}, Fig 6(b)	18.0533
Peppers.jpg, Fig 6(a)	2 points of insertion , P21{inf, sup}, Fig 6(c)	18.0533
Peppers.jpg, Fig 6(a)	3 points of insertion, P21{inf, sup} & P22[sup}, Fig 6(d)	17.7247

PSNR is Peak Signal to Noise Ratio between original image and watermarked image is given by

$$PSNR = 10 \log_{10} (X^2_{max} / MSE) \quad (1)$$

Where X_{max} : is maximum luminance. B bit per sample has X_{max} equal to 2^B-1.

MSE is mean-square-error between original image and watermarked image given by

$$MSE = \sum_{i=1}^{N} \sum_{j=1}^{M} (I_{ij} - J_{ij})^2 / NM \quad (2)$$

Where N X M pixels of original image I and watermarked Image J.

VIII. CONCLUSION AND FUTURE SCOPE OF WORK

The digital images are watermarked with encrypted files in order to have invisible watermark. The watermark are encrypted and decrypted to see the image is authentic or it is tried to tamper. The above method is robust as the key used as public key does not lead to any clue for private key. The public key is fuzzy matrix chosen by user and private key is the Fuzzy Rough set for embedded watermark. The four or more parts of files can embedded into image in the respective region using appropriate Fuzzy Rough set to get desired results. It can tolerate attacks like compression, geometric filters and noise filters. The watermark is robust against changes in file format. These embedding methods can be used for all file formats. The watermark is extracted and decrypted using private key by other party. Further, the work is to be extended for achieving robustness and restraining many more types of attacks.

REFERENCES

[1] A. Menezes, P. van Oorschot, and S. Vanstone, "Handbook of Applied Cryptography", CRC Press, 1996.

[2] An Introduction to Cryptography, 1999 Network Associates, Inc. and its Affiliated Companies.

[3] Glenn Durfee, PhD thesis "Cryptanalysis Of RSA Using Algebraic And Lattice Methods", June 2002, Stanford University.

[4] Alina Mihaela Oprea ,"Efficient Cryptographic Techniques for Securing Storage Systems" CMU-CS-07-119April 2007, PhD thesis, School of Computer Science Carnegie Mellon University ,Pittsburgh, PA 15213.

[5] Shafi Goldwasser, Mihir Bellare,"Lecture Notes on Cryptography", MIT,July 2008.

[6] Whitefiled Diffie and Martin E. Hellman , "Multiuser Cryptographic Techniques*" , *Stanford University* Stanford, California.

[7] Kenneth W. Dam and Herbert S. Lin, **"**Cryptography's Role in Securing the Information Society", Editors, Committee to Study National Cryptography Policy, National Research Council.

[8] Harvey Cohn, "Advanced Number Theory", Distinguished Professor, City University New York.

[9] Rethinking Public Key Infrastructures and Digital Certificates, Rethinking Public Key Infrastructures and Digital Certificates, MIT PressR.

[10] Benny Pinkas, "Cryptographic Techniques for Privacy Preserving Data Mining", HP Labs, benny.pinkas@hp.com

[11] Sharbani Bhattacharya , "Data Security :Issue in Cloud Computing for e-Learning" , Published in University School of Management Studies ,Guru Govind Singh Indraprastha University in National Conference on Information Management 20th March 2010 on. Conference proceedings & Book in – "Information Management in Knowledge Economy", MacMillan Publishers Page 165.

[12] Xinyuan Wang and Douglas S. Reeves "Robust Correlation of Encrypted Attack Traffic Through Stepping Stones By Watermarking The Interpacket Timing", Dept. of Computer Science ,North Carolina State University, Raleigh, NC 27695.

[13] E. C. C. Tsang, Changzhong Wang, Degang Chen, Congxin Wu and Qinghua Hu, "Communication Between Information Systems Using Fuzzy Rough Sets", IEEE Transactions On Fuzzy Systems, Vol. 21, No. 3, June 2013, Page 527.

[14] Teresa Garcia-Valverde, Alberto Garcia-Sola, Hani Hagras, James A. Dooley,Victor Callaghan and Juan A. Botia, "A Fuzzy Logic-Based System for Indoor Localization Using WiFi in Ambient Intelligent Environments", IEEE Transactions On Fuzzy System, Vol. 21, No. 4, August 2013, Page 702.

[15] Hooman Tahayori, Alireza Sadeghian *and* Witold Pedrycz, "Induction of Shadowed Sets Based on the Gradual Grade of Fuzziness", ", IEEE Transactions On Fuzzy System, Vol. 21, No. 5, October 2013,Page 937.

[16] S.P.Tiwari and Arun K.Srivastava,"Fuzzy rough sets,fuzzy preorders and fuzzy topologies", Elsevier, Fuzzy Sets and Systems, Vol. 210 ,2013,Page 63.

[17] Dong Zheng, Yan Liu, Jiyang Zhao, and Abdulmotelab El Sadikk **, "**A Survey of RST Invariant Image Watermarking Algorithms" **, ,***University of Ottawa* ACM Computing Surveys", Vol. 39, No. 2, Article 5, Publication date: June 2007.

[18] Sameh Oueslati, Adnane Cherif and Basel Solaiman, "A Fuzzy Watermarking Approach Based on the Human Visual System", page 218, International Journal Of Image Processing, Volume (4): Issue (3). 2010.

[19] Anderson Rocha, Walter Scheirer and Terrance Boult, Siome Goldenstein, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics" ACM Computing Surveys, Vol. 43, No. 4, Article 26, Publication date: October 2011.

[20] Chuntao WangJiangqun Ni,and Jiwu, "An Informed Watermarking Scheme Using Hidden Markov Model in the Wavelet Domain" Huang,IEEE Transactions On Information Forensics And Security, Vol. 7, No. 3, June 2012 , Page853.

[21] J. C. Kelkboom, Jeroen Breebaart, Ileana Buhan, and Raymond N. J.Veldhuis, "Maximum Key Size and Classification Performance of Fuzzy Commitment for Gaussian Modeled Biometric Sources" IEEE Transaction On Information Forensic and Security, Vol. 7, No. 4, August 2012 Page No. 122.

[22] Peyman Rahmati, Andy Adler and Thomas Tran , " Watermarking in E-commerce", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, 2013.

[23] T.Sridevi and S Sameena Fatima,"Digital Image Watermarking using Fuzzy Logic approach based on DWT and SVD", International Journal of Computer Applications ,Vol 74– No.13, July 2013.

[24] Mriganka Gogoi , H.M.Khalid Raihan Bhuyan ,Koushik Mahanta , Dibya Jyoti Das and Ankita Dutta, " Image and Video based double watermark extraction spread spectrum watermarking in low variance region", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, July 2013.

[25] Sharbani Bhattacharya, "Encrypting Watermark by Fuzzy Max-Min Matrix" presented in National Conference on Advances in Mobile Communications, Networking and Computing, MCNC 2013, ICEIT Conference on 27 -28 September 2013 at IIC, Lodhi Road, New Delhi published in Conference Proceedings, Page 150.

[26] Alimohammad Latif, " An Adaptive Digital Image Watermarking Scheme using Fuzzy Logic and Tabu Search", Journal of Information Hiding and Multimedia Signal Processing, Volume 4, Number 4, October 2013.

[27] Elzbieta Zielinska, Wojciech Mazurczyk and Krzysztof Szczypiorski, "Trends in Steganography", Communication of ACM, Vol 57 ,No. 3, March 2014.

[28] Qian Ying , Ren Xue-mei, Huang Ying and Meng Li, " Image Sharpness Metric Based on Algebraic Multi- Grid Method",International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4,April 2014.

[29] Sharbani Bhattacharya, "Watermarking Digital Image Using Fuzzy Matrix Rules", presented in National Conference Smarter Approaches in Computing , Technology & Applications SACTA 2014 at ITS ,Mohan Nagar, Ghaziabad 19th April 2014, published in Conference Proceedings Page No 343.

The Solution Structure and Error Estimation for The Generalized Linear Complementarity Problem

Tingfa Yan
Middle school of Xiao Bu Ling
Tancheng, Shandong, 276134 P.R. China

***Abstract*—In this paper, we consider the generalized linear complementarity problem (GLCP). Firstly, we develop some equivalent reformulations of the problem under milder conditions, and then characterize the solution of the GLCP. Secondly, we also establish the global error estimation for the GLCP by weakening the assumption. These results obtained in this paper can be taken as an extension for the classical linear complementarity problems.**

Keywords—GLCP; solution structure; error estimation

I. INTRODUCTION

Let mappings $F(x) = Mx + p$, $G(x) = Nx + q$. The generalized linear complementarity problem, abbreviated as GLCP, is to find vector $x^* \in R^n$ such that

$$F(x^*) \in K, G(x^*) \in K^0, F(x^*)^T G(x^*) = 0, \quad (1)$$

where M, $N \in R^{m\times n}$, $p, q \in R^m$, K is a polyhedral cone in R^m, that is, there exist $A \in R^{s\times m}$, $B \in R^{t\times m}$, such that

$$K = \{v \in R^m \mid Av \geq 0, Bv = 0\}.$$

It is easy to verify that its polar cone K^0 assumes the following from

$$K^0 = \{u \in R^m \mid u = A^T\lambda_1 + B^T\lambda_2, \lambda_1 \in R^s_+, \lambda_2 \in R^t\}.$$

The solution set of the GLCP is denoted by X^*, which is assumed to be nonempty throughout this paper.

The GLCP is a direct generalization of the classical linear complementarity problem (LCP) and a special case of the general variational inequalities problem(GVI)([1]). The GLCP was deeply discussed [1, 2, 3, 4] after the work in [5]. The GLCP plays a significant role in economics, engineering, supply chain network equilibrium, etc. ([6, 7, 8, 9]). For example, the balance of supply and demand is central to all economic systems; mathematically, this fundamental equation in economics is often described by a complementarity relation between two sets of decision variables ([9]). Furthermore, the classical Walrasian law of competitive equilibria of exchange economies can be formulated as a generalized nonlinear complementarity problem in the price and excess demand variables ([7]). At the same time, the GLCP be also found applications in contact mechanics problems, structural mechanics problems, obstacle problems mathematical physics, traffic equilibrium problems, etc ([9]), and has been received much attention of researchers.

Up to now, the issues of solution structure and numerical solution methods for GLCP were fully discussed in the literature (e.g., [2, 3, 4, 5, 10, 11, 12]). To our knowledge, Mangasarian and Shiau ([13]) are the first one who gave the solution structure and error estimation analysis to LCP. Latter, Mathias and Pang ([14]) established the solution structure and global error estimation for the LCP with a P-matrix in terms of the natural residual function, and Mangasarian and Ren gave the same error estimation of the LCP with an R_0 -matrix in [15].

Using the implicit Lagrangian function, Luo et al. ([16]) established a global error estimation for the LCP with a nondegenerate matrix. Obviously, the GLCP is an extension of the LCP, the following questions are posed naturally: How about the error estimation for the GLCP? Can the existing error estimation for the LCP be extended to the GLCP? Thus, this motivates us to extend the solution structure and error estimation conclusions of the LCP to the GLCP.

On the other hand, the error estimation for the GLCP was also fully analyzed (e.g., [4, 10, 11]). This paper is a follow-up to [4, 10], as in these papers will establish the global error estimation of the GLCP under weaker conditions than that needed in [4].

The paper is organized as follows. In Section 2, we present some equivalent reformulations of the problem under milder conditions, and detect the solution characterization of the GLCP under milder conditions. The global error estimation is also established for the GLCP in Section 3. Section 4 concludes this paper. These constitute which can be taken as extensions of those for LCP.

Some notations used in this paper are in order. Vectors considered in this paper are all taken in Euclidean space R^n equipped with the standard inner product. The 2-norm of vector in the space is denoted by $\|\cdot\|$. We use R^n_+ to denote the nonnegative orthant in R^n, use x_+ and x_- to denote the vectors composed by elements

$$(x_+)_i := \max\{x_i, 0\}, (x_-)_i := \max\{-x_i, 0\}, 1 \leq i \leq n,$$

respectively, and use $\mathrm{dist}(x, X^*)$ to denote the distance from a point x to the solution set X^*. For simplicity, we also use $x \geq 0$ to denote a nonnegative vector $x \in R^n$ if there is no confusion.

II. The Solution Structure for GLCP

In this section, we mainly present the characterization of the solution for GLCP. First, we give the needed assumptions and some known results from [4] for GLCP.

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.

Remark 1. Under Assumption(A2), A^T has full-column rank and it has left inverse $(AA^{\bullet})^{-1}A$, which is also its pesudo-inverse of A^T. On the other hand, the condition that the matrix A^T has full-column rank is weaker than that the matrix $(A^{\bullet}, B^{\bullet})$ has full-column rank discussed in [4].

Under Assumption (A2), we can establish the following equivalent formulation of the GLCP([4]). i.e., x is a solution of the GLCP if and only if x is a solution of the following system

$$\begin{cases} AF(x) \geq 0, \\ BF(x) = 0, \\ (F(x))^{\bullet}\, G(x) = 0, \\ UG(x) \geq 0, \\ VG(x) = 0, \end{cases} \tag{2}$$

where

$$U = \{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\},$$
$$V = \{A^{\bullet}\,\{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\}$$
$$+B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] - I\}.$$

From (2), using the first equality and the last equality, and combining the first and the second inequality in (2), for any $x \in R^n$, we can obtain

$$\begin{aligned} &(F(x))^{\bullet}\, G(x) \\ &= (F(x))^{\bullet}\,\{A^{\bullet}\,\{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+} \\ &\times[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\} \\ &+B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I]\}G(x) \\ &= [AF(x)]^{\bullet}\,\{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+} \\ &\times[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\}G(x) \\ &+[BF(x)]^{\bullet}\,\{[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I]\}G(x) \\ &= [AF(x)]^{\bullet}\,[UG(x)] \geq 0. \end{aligned} \tag{3}$$

Thus, system (2) can be further written as

$$\begin{cases} AF(x) \geq 0, \\ BF(x) = 0, \\ (AF(x))^{\bullet}\,[UG(x)] = 0, \\ UG(x) \geq 0, \\ VG(x) = 0. \end{cases} \tag{4}$$

Combining (2) with (3), we can establish the following optimization reformulation of the GLCP, and one has that x^* is a solution of the GLCP if and only if x^* is its global optimal solution with the objective vanishing:

$$\begin{aligned} \min \quad & H(x) = (Mx + p)^{\bullet}\,(Nx + q) \\ \text{s.t.} \quad & x \in \Omega, \end{aligned} \tag{5}$$

where $\Omega = \left\{ x \in R^n \,\middle|\, AF(x) \geq 0, BF(x) = 0, UG(x) \geq 0, VG(x) = 0 \right\}$.

Under Assumption (A1), $H(x)$ is a convex function, and Ω is also a convex set. Thus, (5) is a standard convex optimization, we know that its solution set coincides with its stationary point set, i.e., the solution set of the following variational inequality problem: find $x^* \in \Omega$ such that

$$(x - x^*)^{\bullet}\,(\bar{M}x^* + \bar{q}) \geq 0, \forall x \in \Omega, \tag{6}$$

where $\bar{M} = M^{\bullet}\,N + N^{\bullet}\,M$, $\bar{q} = M^{\bullet}\,q + N^{\bullet}\,p$.

Theorem 1 Suppose that Assumption (A1) and (A2) hold, and x_0 is a solution of the GLCP. Then

$$X^* = \left\{ x \in X \,\middle|\, \bar{M}(x - x_0) = 0, (x - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) = 0 \right\}.$$

Proof. Set

$$W = \left\{ w \in X \,\middle|\, \bar{M}(w - x_0) = 0, (w - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) = 0 \right\}.$$

For any $\tilde{x} \in X^*$, since $\tilde{x}, x_0 \in X$, by (6), we get

$$(\tilde{x} - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) \geq 0, \tag{7}$$
$$(x_0 - \tilde{x})^{\bullet}\,(\bar{M}\tilde{x} + \bar{q}) \geq 0. \tag{8}$$

Combining (7) with (8), we can obtain

$$(\tilde{x} - x_0)^{\bullet}\,\bar{M}(\tilde{x} - x_0) \leq 0.$$

Combining this with Assumption (A1), we get

$$(\tilde{x} - x_0)^{\bullet}\,\bar{M}(\tilde{x} - x_0) = 0, \tag{9}$$

and conclude that

$$\bar{M}(\tilde{x} - x_0) = 0. \tag{10}$$

Combining (9) and (8), one has

$$\begin{aligned} (x_0 - \tilde{x})^{\bullet}\,(\bar{M}x_0 + \bar{q}) &= (x_0 - \tilde{x})^{\bullet}\,\bar{M}(x_0 - \tilde{x}) \\ &+ (x_0 - \tilde{x})^{\bullet}\,(\bar{M}\tilde{x} + \bar{q}) \geq 0. \end{aligned}$$

Combining this with (7), we have

$$(\tilde{x} - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) = 0.$$

Combining this with (10), we obtain $\tilde{x} \in W$.

On the other hand, for any $w \in W$, since

$$\bar{M}(w - x_0) = 0, (w - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) = 0,$$

for any $x \in X$, using the fact that x_0 is a solution of the GLCP, combining (6), we obtain

$$\begin{aligned} &(x - w)^{\bullet}\,(\bar{M}w + \bar{q}) \\ &= [(x - x_0) - (w - x_0)]^{\bullet}\,[\bar{M}(w - x_0) + (\bar{M}x_0 + \bar{q})] \\ &= [(x - x_0) - (w - x_0)]^{\bullet}\,[(\bar{M}x_0 + \bar{q})] \\ &= (x - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) - (w - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) \\ &= (x - x_0)^{\bullet}\,(\bar{M}x_0 + \bar{q}) \geq 0, \end{aligned}$$

Thus, $w \in X^*$.

Theorem 2 If x_1 and x_2 are two solutions of the GLCP. Then, $(Mx_1 + p)^{\bullet}\,(Nx_2 + q) = (Mx_2 + p)^{\bullet}\,(Nx_1 + q) = 0$.

Proof. Since x_1 and x_2 are two solutions of the GLCP. By Theorem 1, we have

$$\bar{M}(x_1 - x_2) = \bar{M}(x_1 - x_0) - \bar{M}(x_2 - x_0) = 0.$$

Combining this with the fact that x_1 and x_2 are two solutions of the GLCP, we have

$$(Mx_1 + p)^{\cdot}(Nx_1 + q) = 0, (Mx_2 + p)^{\cdot}(Nx_2 + q) = 0,$$

one has

$$\begin{aligned} 0 &= (x_1 - x_2)^{\cdot}(M^{\cdot}N + N^{\cdot}M)(x_1 - x_2) \\ &= 2(x_1 - x_2)^{\cdot}M^{\cdot}N(x_1 - x_2) \\ &= 2[(Mx_1 + p) - (Mx_2 + p)]^{\cdot}[(Nx_1 + q) - (Nx_2 + q)] \\ &= -2[(Mx_1 + p)^{\cdot}(Nx_2 + q) + (Mx_2 + p)^{\cdot}(Nx_1 + q)]. \end{aligned} \tag{11}$$

Using the similar technique to that of (3) , we can deduce

$$(Mx_1 + p)^{\cdot}(Nx_2 + q) \geq 0, (Mx_2 + p)^{\cdot}(Nx_1 + q) \geq 0. \tag{12}$$

Combining (11) with (12), and the desired result follows.

Theorem 3 The solution set of GLCP is a convex set.

Proof. If solution set of the GLCP is single point set, then it is obviously convex. In this following, we suppose that x_1 and x_2 are two solutions of the GLCP. By Theorem 1, we have

$$\begin{aligned} &\bar{M}(x_1 - x_0) = 0, \bar{M}(x_2 - x_0) = 0, \\ &(\bar{M}x_0 + \bar{q})^T(x_1 - x_0) = 0, \\ &(\bar{M}x_0 + \bar{q})^T(x_2 - x_0) = 0. \end{aligned} \tag{13}$$

For vector $x = \tau x_1 + (1-\tau)x_2, \forall\, \tau \in [0,1]$, by (13), we have

$$\begin{aligned} \bar{M}(x - x_0) &= \bar{M}[\tau x_1 + (1-\tau)x_2 - x_0] \\ &= \tau\bar{M}(x_1 - x_0) + (1-\tau)\bar{M}(x_2 - x_0) \\ &= 0. \end{aligned} \tag{14}$$

$$\begin{aligned} &(\bar{M}x_0 + \bar{q})^T(x - x_0) \\ &= (\bar{M}x_0 + \bar{q})^T[\tau x_1 + (1-\tau)x_2 - x_0] \\ &= \tau(\bar{M}x_0 + \bar{q})^T(x_1 - x_0) \\ &\quad + (1-\tau)(\bar{M}x_0 + \bar{q})^T(x_2 - x_0) = 0. \end{aligned} \tag{15}$$

Combining (14), (15) with the conclusion of Theorem 1, we obtain the desired result.

III. The Error Estimation for GLCP

In this section, we will present a global error estimation for the GLCP under weaker conditions than that needed in [4]. Firstly, we can give the needed error bound for a polyhedral cone from [17] and error bound for a convex optimization from [18] to reach our aims.

Lemma 1 For polyhedral cone

$$P = \{x \in R^n \mid D_1 x = d_1, B_1 x \leq b_1\}$$

with $D_1 \in R^{l\times n}$, $B_1 \in R^{m\times n}$, $d_1 \in R^l$ and $b_1 \in R^m$, there exists a constant $c_1 > 0$ such that

$$\operatorname{dist}(x, P) \leq c_1[\| D_1 x - d_1\| + \|(B_1 x - b_1)_+\|], \quad \forall x \in R^n.$$

Lemma 2 Let P be a convex polyhedron in R^n and θ be a convex quadratic function defined on R^n. Let S be the nonempty set of globally optimal solutions of the program:

$$\begin{aligned} \min \quad & \theta(x) \\ \text{s.t.} \quad & x \in P \end{aligned}$$

with θ_{opt} being the optimal value of θ on S. There exists a scalar $c_2 > 0$ such that

$$\begin{aligned} \operatorname{dist}(x, S) \leq c_2 \max\{&\operatorname{dist}(x, P), |[\theta(x) - \theta_{opt}]_+|, \\ &|[\theta(x) - \theta_{opt}]_+|^{(1/2)}\}, \forall x \in R^n. \end{aligned}$$

Theorem 4 Under Assumption 2.1 (A1) and (A2), then there exists constant $\rho > 0$ such that

$$\begin{aligned} \operatorname{dist}(x, X^*) \leq \rho\Big\{&\|[AF(x)]_-\| + \|BF(x)\| + \|[UG(x)]_-\| \\ &+ \|VG(x)\| + |[F(x)^{\cdot}G(x)]_+| \\ &+ |[F(x)^{\cdot}G(x)]_+|^{(1/2)}\Big\}, \forall x \in R^n. \end{aligned}$$

Proof. For problem (5), under Assumption 2.1 (A1), $H(x)$ is a convex function, and we know that x^* is a solution of the GLCP if and only if x^* is its global optimal solution with the objective vanishing, i.e., $H(x)_{opt} = 0$. For any $x \in R^n$, a direct computation yields that

$$\begin{aligned} \operatorname{dist}(x, X^*) &\leq c_3 \max\{\operatorname{dist}(x, \Omega), |[H(x)]_+|, |[H(x)]_+|^{(1/2)}\} \\ &\leq c_3 \max\{c_4\{\|[AF(x)]_-\| + \|BF(x)\| + \|[UG(x)]_-\| \\ &\quad + \|VG(x)\|\}, |[H(x)]_+|, |[H(x)]_+|^{(1/2)}\} \\ &\leq c_3 \max\{c_4, 1\}\{\|[AF(x)]_-\| + \|BF(x)\| \\ &\quad + \|[UG(x)]_-\| + \|VG(x)\| \\ &\quad + |[(F(x))^{\cdot}G(x)]_+| + |[(F(x))^{\cdot}G(x)]_+|^{(1/2)}\}, \end{aligned} \tag{16}$$

where the first inequality follows from Lemma 2 with constant $c_3 > 0$, and the second inequality uses Lemma 1 with constant $c_4 > 0$. Using (16), letting $\rho = c_3 \max\{c_4, 1\}$, the desired result follows.

Remark 2. Combining Remark 1. Assumption 1 (A2) in Theorem 4 is weaker than the Assumption (A2) in Theorem 4.1 in [4], and the Assumption (A1) in this paper coincides with Assumption (A1) in [4]. In addition, Theorem 4 is sharper than Theorem 4.1 in [4].

In the end of this paper, we will consider a special case of GLCP which was discussed in [13, 14, 15, 16].

When $K = R^n_+, F(x) = x$, then $K^0 = R^n_+$, and GLCP reduces to the LCP of finding vector $x^* \in R^n$ such that

$$x^* \geq 0, Nx^* + q \geq 0, (x^*)^T(Nx^* + q) = 0. \tag{17}$$

Combining (17) with Theorem 4, we can immediately obtain the following conclusion.

Corollary 1 Suppose the matrix N is semi-definite (not necessarily symmetric), and $F(x) = x$. Then there exists constant $\rho_1 > 0$ such that

$$\operatorname{dist}(x, X^*) \le \rho_1 \Big\{ \| x_- \| + \| (Nx+q)_- \| + | [x^{\bullet} (Nx+q)]_+ | + | [x^{\bullet} (Nx+q)]_+ |^{(1/2)} \Big\}, \forall x \in R^n.$$

Proof. By $K = R^n_+$, we have $A = I, B = 0$, from $F(x) = x$, we have $M = I, q = 0$, where I is an identity matrix. Combining this with definition of U, V in (2), we can obtain $U = I, V = 0$. Combining these with Theorem 4, then the desired result follows.

Remark 3. It is clear that the assumption in Corollary 1 above coincides with that in Theorem 2.7 in [13]. Furthermore, the conclusion in Corollary 1 is stronger than that in Theorem 2.7 in [13].

When $K = R^n_+$, GLCP reduces to the vertical linear complementarity problem([19])of finding vector $x^* \in R^n$ such that

$$Mx^* + q \ge 0, Nx^* + q \ge 0, (Mx^* + q)^T (Nx^* + q) = 0. \qquad (18)$$

Combining (18) with Theorem 4, we have the following conclusion hold.

Corollary 2 Under Assumption 2.1 (A1). Then there exists constant $\rho_2 > 0$ such that

$$\operatorname{dist}(x, X^*) \le \rho_2 \Big\{ \| (Mx+p)_- \| + \| (Nx+q)_- \| + | [(Mx+p)^{\bullet} (Nx+q)]_+ | + | [(Mx+p)^{\bullet} (Nx+q)]_+ |^{(1/2)} \Big\}, \forall x \in R^n.$$

Proof. By $K = R^n_+$, one has $A = I, B = 0$. Combining this with definition of U, V in (2), one has $U = I, V = 0.$ Combining these with Theorem 4, then the desired result follows.

Remark 4. Obviously, Assumption 2.1(A1) in Corollary 2 above is weaker than that in Corollary 2 in [19], since the condition which rank $\left(M^T, N^T\right)^T = n$ is removed.

IV. Conclusions

In this paper, we presented the solution characterization for GLCP, and established global error estimation on the GLCP under weaker conditions than that needed in [4], which is the extension of this for LCP. Surely, under milder conditions, we may also established the solution structure and error estimation for GLCP such as the mapping being nonmonotone involved in the GLCP, this is a topic for future research.

Acknowledgment

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 (ZR2011FL017).

References

[1] M.A. Noor, “General variational inequalities”, Appl. Math. Letters, 1(2), pp. 119-121, 1988.

[2] Y.J. Wang, F.M. Ma, J.Z. Zhang, “A nonsmooth L-M method for solving the generalized nonlinear complementarity problem over a polyhedral cone”, Appl. Math. Optim., 52, pp. 73-92, 2005.

[3] X.Z. Zhang, F.M. Ma, Y.J. Wang, “A Newton-type algorithm for generalized linear complementarity problem over a polyhedral cone”, Appl. Math. Comput., 169, pp. 388-401, 2005.

[4] 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),pp.93-107, 2013.

[5] R. Andreani, A. Friedlander, S.A. Santos, “On the resolution of the generalized nonlinear complementarity problem”, SIAM J. Optim., 12, pp. 303-321, 2001.

[6] F. Facchinei and J.S. Pang , Finite-dimensional variational inequality and complementarity problems, Springer, New York, NY, 2003.

[7] L. Walras , Elements of pure economics, Allen and Unwin, London, 1954.

[8] L.P. Zhang, “A nonlinear complementarity model for supply chain network equilibrium”, Journal of Industrial and Managment Optimization, 3(4), pp.727-737, 2007.

[9] M.C. Ferris, J.S. Pang, “Engineering and economic applications of complementarity problems”, Society for industrial and applied mathematics, 39(4), pp.669-713, 1997.

[10] Y. Z. Diao, “An error estimation for management equilibrium model”, International Journal of Computer and Information Technology, 2(4), pp.677-681,2013.

[11] S.S. Xie, P. Wang, L. Wang, H.C. Sun, “An algorithm for the nonlinear complementarity problem on management equilibrium model”, International Journal of Computer and Information Technology, 2(6), pp.1136-1140, 2013.

[12] L. Wang, “A global convergence algorithm for the supply chain network equilibrium model”, International Journal of Advanced Computer Science and Applications, 3(2), pp.15-18,2012.

[13] O.L. Mangasarian and T.H. Shiau, “Error bounds for monotone linear complementarity problems”, Math. Programming, 36(1): 81--89, 1986.

[14] R. Mathias and J.S. Pang, “Error bound for the linear complementarity problem with a P-matrix”, Linear Algebra & Appl., 132: 123-136, 1990.

[15] O.L. Mangasarian and J. Ren, “New improved error bound for the linear complementtarity problem”, Math. Programming, 66: 241-255, 1994.

[16] Z.Q. Luo, O.L. Mangasarian, J. Ren and M.V. Solodov, “ New error bound for the linear complementarity problem”, Math. Operations Research, 19: 880-892, 1994.

[17] A.J. Hoffman, “On the approximate solutions of linear inequalities”, Journal of Research of the National Bureau of Standards, 49, pp. 263-265, 1952.

[18] T. Wang and J.S. Pang, “Global error bounds for convex quadratic inequality systems”, Optimization, 31, pp. 1-12, 1994.

[19] J.Z. Zhang, N.H. Xiu, “Global s-type error bound for the extended linear complementarity problem and applications” , Math. Program., Ser. B, 88: 391- 410, 2000.

Forecasting Rainfall Time Series with stochastic output approximated by neural networks Bayesian approach

Cristian Rodriguez Rivero
Department of Electronic Engineering
Universidad Nacional de Córdoba
Córdoba, Argentina

Julian Antonio Pucheta
Department of Electronic Engineering
Universidad Nacional de Córdoba
Córdoba, Argentina

***Abstract*— The annual estimate of the availability of the amount of water for the agricultural sector has become a lifetime in places where rainfall is scarce, as is the case of northwestern Argentina. This work proposes to model and simulate monthly rainfall time series from one geographical location of Catamarca, Valle El Viejo Portezuelo. In this sense, the time series prediction is mathematical and computational modelling series provided by monthly cumulative rainfall, which has stochastic output approximated by neural networks Bayesian approach. We propose to use an algorithm based on artificial neural networks (ANNs) using the Bayesian inference. The result of the prediction consists of 20% of the provided data consisting of 2000 to 2010. A new analysis for modelling, simulation and computational prediction of cumulative rainfall from one geographical location is well presented. They are used as data information, only the historical time series of daily flows measured in mmH_2O. Preliminary results of the annual forecast in mmH_2O with a prediction horizon of one year and a half are presented, 18 months, respectively. The methodology employs artificial neural network based tools, statistical analysis and computer to complete the missing information and knowledge of the qualitative and quantitative behavior. They also show some preliminary results with different prediction horizons of the proposed filter and its comparison with the performance Gaussian process filter used in the literature.**

Keywords—rainfall time series; stochastic method; bayesian approach; computational intelligence

I. INTRODUCTION

Climate variability in the semi-humid and arid parts of the northwestern part of Argentina poses a great risk to the people and resources of these regions [1] as the smallest fluctuations of weather parameters like precipitation not only damage the agriculture and economy of the region but disturb the overall water cycle [2].

The ANNs are mostly used as predictor filter with an unknown number of parameters performed by a lot of author, recently, such as in [3][4][5][6]. One famous black box model that forecast rainfall time series in recent decades is artificial neural network model. Artificial neural networks are free-intelligent dynamic systems models that are based on the experimental data, and the knowledge and covered law beyond data changes to network structure by trends on these data [7]. The difficulties in modeling such complex systems are considerably reduced by the recent Artificial Intelligence tools like Artificial Neural Networks (ANNs); Genetic Algorithm (GA) [8] based evolutionary optimizer and Genetic Programming (GP).

In turn, this work propose to estimate water availability horizon useful for control problems in agricultural activities such as seedling growth and decision-making using some ANNs approaches presented in recent earlier works [9]. An ANNs filter is used and their parameters are set in function of the roughness of the time series. These are considered as random variables whose distribution is inferred by posterior probability from the data, in which is included as an additional parameter, the number of hidden neurons and modelling uncertainty [10].

The Bayesian approach permits propagation of uncertainty in quantities which are unknown to other assumptions in the model, which may be more generally valid or easier to guess in the problem. For neural networks, the Bayesian approach was pioneered in [11]-[12], and reviewed [13], [14] and [15]. The main difficulty in model building is controlling the complexity of the model. It is well known that the optimal number of degrees of freedom in the model depends on the number of training samples, amount of noise in the samples and the complexity of the underlying function being estimated.

The procedure of determining the prior density and likelihood functions associated with rainfall time series uncertainty is very complicated and there is a requirement to assume a linear and normal distribution within the framework of the proposed parameters. The problem of model selection is often divided into discover an organization of a model's parameters that is well-matched such as the network topology, e.g. number of patterns, layers, hidden units per layer, that results in the best generalization performance. A common result is with too many free parameters tend to overfit the training data and, thus, show poor generalization performance.

A model attempting to estimate the value of a random variable may have potential access to a wide range of measurements regarding the state of the environment. Some of these quantities may provide the model with useful information regarding the random variable, whereas others may not. In the context of neural networks, only the useful quantities should be used as inputs to a network. A network that receives both

useful inputs and "nuisance" inputs will contain too many free parameters and, thus, be prone to overfitting the training data leading to poor generalization.

II. DATA TREATMENT

A rainfall time series can be actually regarded as an integration of stochastic (or random) and deterministic components [16]. Once the stochastic (noise) component is appropriately eliminated, the deterministic component can then be easily modeled. Rainfall is an end product of a number of complex atmospheric processes which vary both in space and time; The data that is available to assist the definition of control variable for the process models, such as rainfall intensity, wind speed, and evaporation, etc. are linked in both the spatial and temporal dimensions; even if the rainfall can be described concisely and completely, the volume of calculations involved may be prohibitive; and the temporal and spatial resolution provided by this approach is not accurate enough for many hydrologic applications. A second approach to forecast rainfall makes use of nonparametric models based on statistics and/or machine learning.

The standard non-parametric approaches presented in this work by means of time-series analysis, is based on stochastic techniques that assume non-linear relationship among data that reproduce the rainfall time series only in statistical sense. Then, in principle, machine learning models, such as artificial neural networks, can improve the forecasting results obtained using models based on standard non-parametric approaches.

The rainfall dataset used is from Cuesta El Portezuelo located at Catamarca, province of Argentina (-28°28'11.26";-65°38'14.05") and the collection date is from year 2000 to 2010 shown in Fig.1.

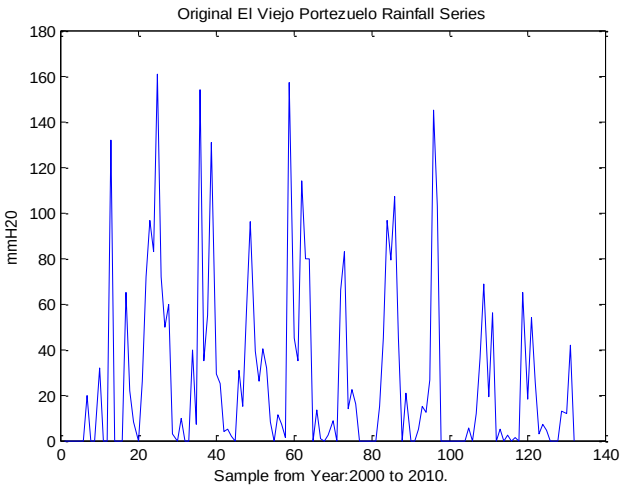


Fig. 1. Original Rainfall times series from El Viejo Portezuelo, Catamarca, Argentina.

III. METHODOLOGY AND BAYESIAN APPROACH

When a time series is being analyzed, it is important to make use of the simplest possible models. Specifically, the number of unknown parameters must be kept at a minimum.

For forecasting problems, Bayesian analysis generates point and interval forecasts by combining all the information and sources of uncertainty into a predictive distribution for the future values. It does so with a function that measures the loss to the forecaster that will result from a particular choice of forecasts.

The gamma distributions have been chosen for this purpose. When a Bayesian analysis is conducted, inferences about the unknown parameters are derived from the posterior distribution. This is a probability model which describes the knowledge gained after observing a set of data. The application of the regression problem involving the correspond neural network function $y(x,w)$ and the data set consisting of N pairs, input vector lx and targets t_n $(n=1,....,N)$.

Assuming Gaussian noise on the target, the likelihood function takes the form:

$$P(D/w,M)=\left(\frac{\beta}{2\pi}\right)^{N/2}\exp\left\{\frac{-\beta}{2}\sum_{n=1}^{N}\left\|y(x_n;w)-t_n\right\|^2\right\}, \quad (8)$$

where β is a hyper-parameter representing the inverse of the noise variance. We consider in this work a single hidden layer of *'tanh'* units and a linear outputs units.

To complete the Bayesian approach for this work, prior information for the network is required. It is proposed to use, analogous to penalties terms, the following equation

$$P(w)=\left(2\pi w^2\right)^{-N/2}\exp\left(\frac{|w|^2}{2w^2}\right), \quad (9)$$

assuming that the expected scale of the weights is given by w set by hand. This was carried out considering that the network function f(xn+1,w) is approximately linear with respect to w in the vicinity of this mode, in fact, the predictive distribution for y_{n+1} will be another multivariate Gaussian.

IV. PROPOSED APPROACH FOR TUNING THE NEURAL NETWORKS BY BAYESIAN APPROACH

In the block diagram of the nonlinear prediction scheme based on a ANN filter is shown. Here, a prediction device [17]-[18] is designed such that starting from a given sequence $\{x_n\}$ at time n corresponding to a time series it can be obtained the best prediction $\{x_e\}$ for the following sequence of 18 values.

Hence, it is proposed a predictor filter with an input vector l_x, which is obtained by applying the delay operator, Z^{-1}, to the sequence $\{x_n\}$. Then, the filter output will generate x_e as the next value, that will be equal to the present value x_n. So, the prediction error at time k can be evaluated as:

$$e(k)=x_n(k)-x_e(k) \quad (6)$$

which is used for the learning rule to adjust the NN weights.

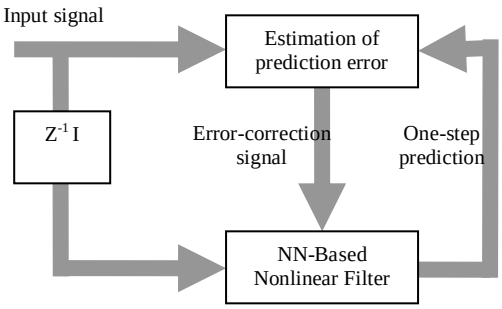


Fig. 2. Block diagram of the nonlinear prediction.

The coefficients of the nonlinear ANNs filter are adjusted on-line in the learning process, by considering an online heuristic criterion that modifies at each pass of the time series the number of patterns, the number of iterations and the length in function of the Hurst's value H calculated from the time series taking into account the Bayesian inference and stochastic dependence of the output values.

A. *Bayesian model*

When a rainfall series is being analyzed, it is important to make use of the simplest possible models. Specifically, the number of unknown parameters must be kept at a minimum. For forecasting problems, Bayesian analysis generates point and interval forecasts by combining all the information and sources of uncertainty into a predictive distribution for the future values. It does so with a function that measures the loss to the forecaster that will result from a particular choice of forecasts.

The gamma distributions have been chosen for this purpose. When a Bayesian analysis is conducted, inferences about the unknown parameters are derived from the posterior distribution. This is a probability model which describes the knowledge gained after observing a set of data. The application of the regression problem involving the correspond neural network function *y(x,w)* and the data set consisting of *N* pairs, input vector lx and targets t_n *(n=1,….,N)*

Assuming Gaussian noise on the target, the likelihood function takes the form:

$$P(D/w,M)=\left(\frac{\beta}{2\pi}\right)^{N/2}\exp\left\{\frac{-\beta}{2}\sum_{n=1}^{N}\left\|y(x_n;w)-t_n\right\|^2\right\}, \qquad (7)$$

where β is a hyper-parameter representing the inverse of the noise variance. We consider in this work a single hidden layer of '*tanh*' units and a linear outputs units. To complete the Bayesian approach for this work, prior information for the network is required. It is proposed to use, analogous to penalties terms, the following equation,

$$P(w)=\left(2\pi w^2\right)^{-N/2}\exp\left(\frac{|w|^2}{2w^2}\right), \qquad (8)$$

assuming that the expected scale of the weights is given by w set by hand. This was carried out considering that the network function f(xn+1,w) is approximately linear with respect to w in the vicinity of this mode, in fact, the predictive distribution for yn+1 will be another multivariate Gaussian.

The computation test results were made on rainfall time series, which consist of 132 data. The Monte Carlo method was employed to forecast the next 18 values with an associated variance. Here it was performed an ensemble of 500 trials with a fractional Gaussian noise sequence of zero mean and variance of 0.11. The fractional noise was generated by the Hosking method [19] with the H parameter estimated from the data time series. The following figures yield the results of the mean and the variance of 500 trials of the forecasted 18 values. Such outcomes for one (30%) and two (69%) sigma are shown in Fig. 6, Fig. 7, Fig. 8, Fig. 10, and Fig. 11. The obtained time series has a mean value, denoted at the foot of the figure by "Forecasted Mean", whereas the "Real Mean" although it is not available at time 114. This procedure is repeated 500 times for each time series.

The assessment of the experimental results has been obtained by comparing the performance of the proposed filter against the Gaussian process based filter. The evolution of the SMAPE index for a neural network bayesian approach filter, which uses a learning algorithm and the GP filter has the same initial parameters in each algorithm, although such parameters and filter's structure are changed by the proposed approach, not is the case of the GP filter. In the proposed filter, the coefficients and the structure of the filter are tuned by considering their stochastic dependency. It can be noted that in each one of Fig. 3 to Fig. 6.

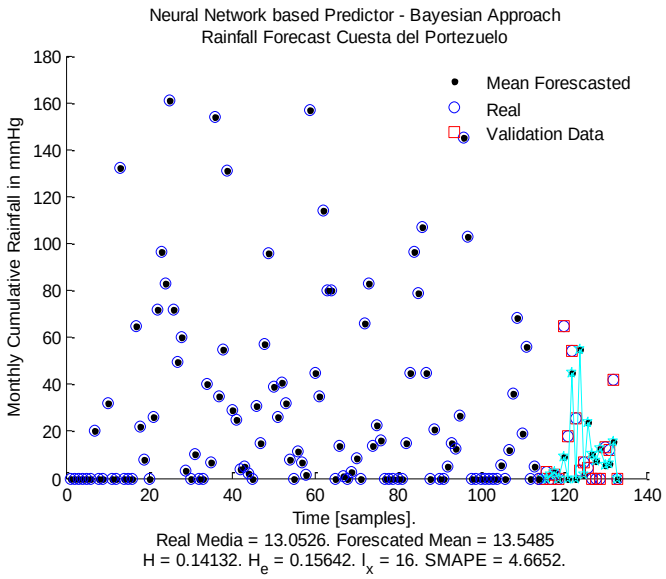


Fig. 3. Cuesta El Portezuelo Rainfall time series neural network Bayesian approach.

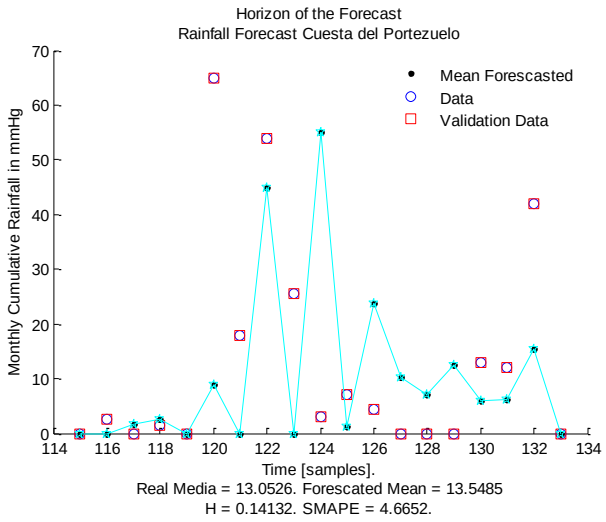


Fig. 4. Cuesta El Portezuelo Rainfall time series Bayesian approach forecast horizon.

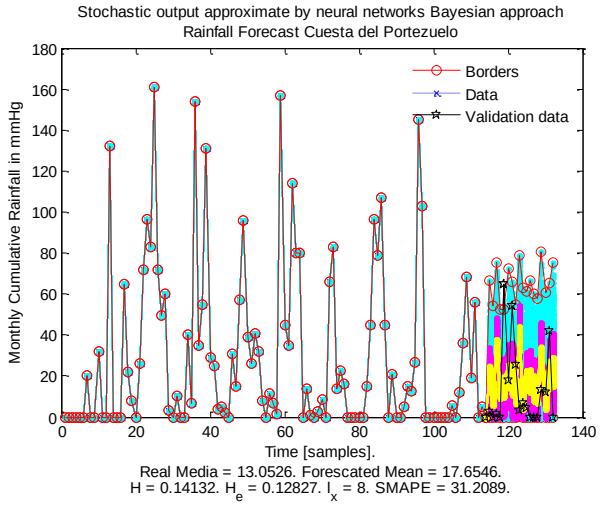


Fig. 5. Cuesta El Portezuelo Rainfall time series stochastic output with zero mean and 0.11 variance.

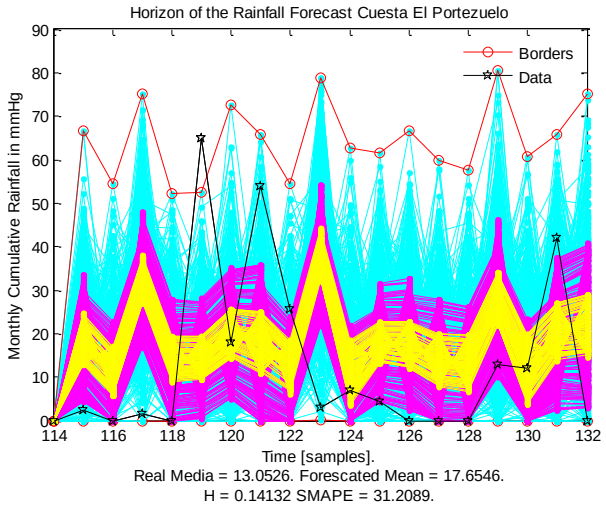


Fig. 6. Cuesta El Portezuelo Rainfall time series stochastic forecast horizon.

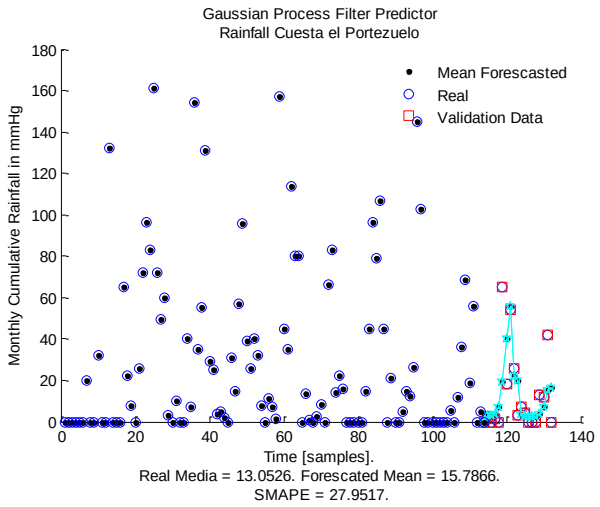


Fig. 7. Cuesta El Portezuelo Rainfall time series Gaussian process filter.

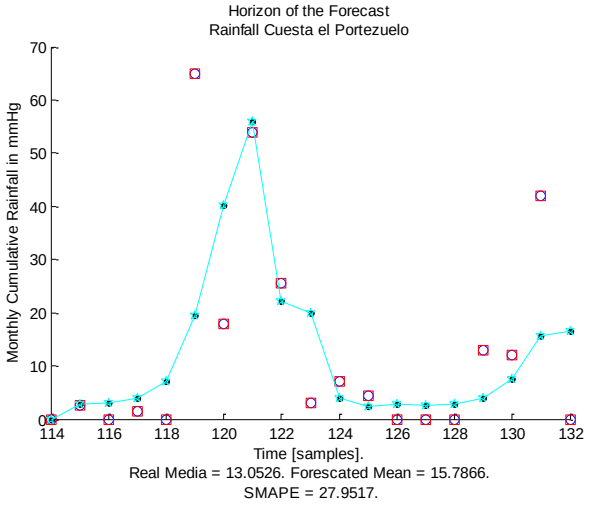
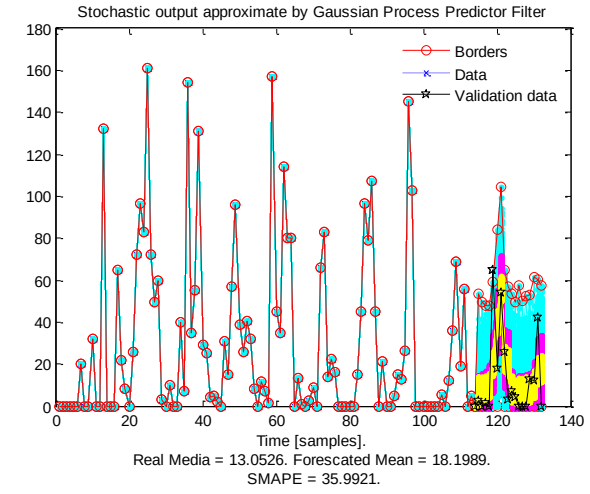


Fig. 8. Cuesta El Portezuelo Rainfall time series Gaussian process filter forecast horizon



.

Fig. 9. Cuesta El Portezuelo Rainfall time series stochastic Gaussian Process output with zero mean and 0.11 variance.

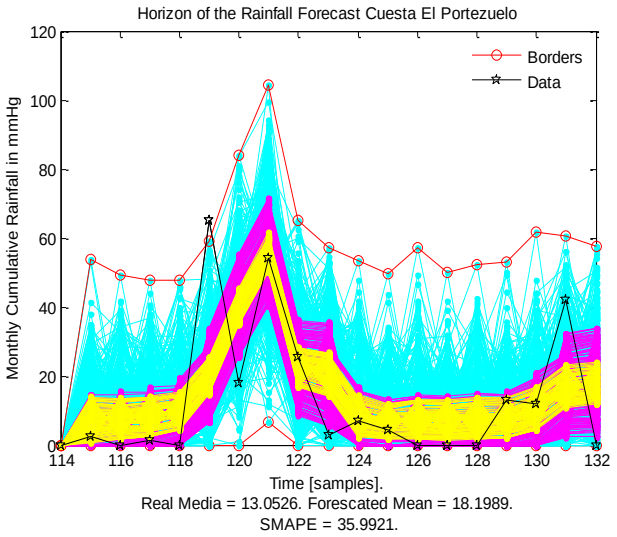


Fig. 10. Cuesta El Portezuelo Rainfall time series stochastic forecast horizon.

The measure of forecast performance is measured by the Symmetric Mean Absolute Percent Error (SMAPE) proposed in the most of metric evaluation, defined by

$$SMAPE_s = \frac{1}{n}\sum_{t=1}^{n}\frac{|X_t - F_t|}{(X_t + F_t + 0.005)/2}\cdot 100 \qquad (9)$$

where t is the observation time, n is the size of the test set, s is each time series, X_t and F_t are the actual and the forecasted time series values at time t respectively. The SMAPE of each series **s** calculates the symmetric absolute error in percent between the actual X_t and its corresponding forecast value F_t, across all observations *t* of the test set of size *n* for each time series **s**.

In each figure are detailed the testing and the computing data, where the testing are labelled "Validation data" and had not been used in the computation of the predictor filter.

In table I, the better performance is shown by the stochastic NN Bayesian approach where the index is set to 4.66 and 31.20. By means of this assessment, the approach can be applied for a class of high roughness rainfall time series, in this case measured by the Hurst parameter [20] to Cuesta El Portezuelo series, H=0.14.

TABLE I. FIGURES OBTAINED BY THE PROPOSED APPROACH FOR EACH PREDICTOR FILTER

Filters for Rainfall time series	Real Mean	Mean Forecasted	SMAPE
NN Bayesian approach	13.05	13.56	4.66
Gaussian Process	13.05	15.78	27.95
Stochastic NN Bayesian approach	13.05	17.65	31.20
Stochastic Gaussian Process	13.05	18.19	35.99

V. CONCLUSIONS

In this article, forecasting rainfall time-series with stochastic output approximated by neural networks Bayesian approach has been presented. In the first case, an ANNs algorithm based on Bayesian inference to model neural networks parameters were detailed. The learning rule proposed to adjust the ANN's coefficients was based on the Levenberg-Marquardt method.

Furthermore, the rainfall series were related with the long or short term stochastic dependence of the time series assessed by the Hurst parameter H, then the stochastic approximation to forecast the next 18 month were implemented. Its main contribution lies in generating stochastic rainfall time series forecast from monthly cumulative rainfall data, which allows adjusting the filter parameters for each algorithm and then averaged over all the outputs. The roughness of the resulting forecasted time series was again evaluated by the Hurst parameter H in the Bayesian approach.

The main results show a good performance of the predictor system based on stochastic neural network Bayesian approach, applied to time series obtained from a geographical point when the observations are taken from a single point due to similar roughness for both, the original and the forecasted time series, respectively.

These results encouraged us to continue working on new machine learning algorithms using novel forecasting methods.

ACKNOWLEDGMENT

This work was supported by Universidad Nacional de Córdoba (UNC), FONCYT-PDFT PRH N°3 (UNC Program RRHH03), SECYT-UNC, Institute of Automatic (INAUT) - Universidad Nacional de San Juan and National Agency for Scientific and Technological Promotion (ANPCyT).

REFERENCES

[1] Center for Studies of Variability and Climate Change (CEVARCAM) - Facultad de Ingeniería y Ciencias Hídricas (FICH), Universidad Nacional del Litoral (UNL).

[2] Magrin, G.raciela, Maria Travasso, Raul Diaz, and Rafael Rodriguez. 1997. "Vulnerability of the agricultural systems of Argentina to climate change", Climate Research 9: 31-36.

[3] Menhaj, M.B., 2012. Artificial Neural Networks. Amirkabir University of Technology.

[4] Cristian M. Rodríguez Rivero, Julián A. Pucheta, Martín R. Herrera, Victor Sauchelli, Sergio Laboret. Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall, (Pronóstico de Series Temporales usando inferencia Bayesiana: aplicación a series de lluvia de agua acumulada). ISSN 1548-0992. Pp. 359 364. IEEE LATIN AMERICA TRANSACTIONS, VOL. 11, NO. 1, FEB. 2013. http://www.ewh.ieee.org/reg/9/etrans/ieee/issues/vol11/vol11issue1Feb.2013/11TLA1_62RodriguezRivero.pdf

[5] Julián A. Pucheta , Cristian M. Rodríguez Rivero , Martín R. Herrera, Carlos A. Salas, H. Victor Sauchelli. "Rainfall Forecasting Using Sub sampling Nonparametric Methods" ("Pronóstico de lluvia usando métodos no paramétricos con submestreo"). ISSN 1548-0992. Pp. 346-350. IEEE LATIN AMERICA TRANSACTIONS, VOL. 11, NO. 1, FEB.2013. http://www.ewh.ieee.org/reg/9/etrans/ieee/issues/vol11/vol11issue1Feb.2013/11TLA1_110Pucheta.pdf

[6] C. Rodriguez Rivero, J. Pucheta, H. Patiño, J. Baumgartner, S. Laboret and V. Sauchelli. "Analysis of a Gaussian Process and Feed-Forward Neural Networks based Filter for Forecasting Short Rainfall Time Series". 2013 International Joint Conference on Neural Networks, Texas, 4 al 9 de Agosto de 2013, USA. Print Edition: IEEE Catalog Number: CENSUS-ART, ISBN: 978-1-4673-6129-3, ISSN: 2161-4407, CD Edition: IEEE Catalog Number: CFPlSUS-CDR, ISBN: 978-1-4673-6128-6. 2013.

[7] Salas, J.D., Delleur, J.W., Yevjevich, V., Lane, W.L. (Eds.), 1985. Applied Modeling of Hydrologic Time Series. Water Resources Publications, Littleton, Colorado.

[8] Baumgartner, J.; Rodríguez Rivero, C. y Pucheta, J. (2011) "Pronóstico de lluvia en un punto desde diversos puntos geográficos de observación mediante procesos gaussianos." XXIII Congreso Nacional del Agua – CONAGUA, 22 al 25 de Junio de 2011, Resistencia, Chaco, Argentina, ISSN 1853-7685.

[9] J. Pucheta, M., C. Rodríguez Rivero, M. Herrera, C. Salas, D. Patiño and B. Kuchen. A Feed-forward Neural Networks-Based Nonlinear Autoregressive Model for Forecasting Time Series. Revista Computación y Sistemas, Centro de Investigación en Computación-IPN, México D.F., México, Computación y Sistemas Vol. 14 No. 4, 2011, pp 423-435, ISSN 1405-5546. http://www.cic.ipn.mx/sitioCIC/images/revista/vol14-4/art07.pdf

[10] C. Rivero Rodríguez, J. Pucheta, J. Baumgartner, M. Herrera, H.D. Patiño and V. Sauchelli, Bayesian modeling of a nonlinear autoregressive filter based on neural networks for monthly cumulative rainfall time series forecasting , anales del Congreso CONAGUA (Congreso Nacional del Agua), ISSN 1853-7685, Chaco, Argentina, realizado del 22 al 25 de Junio, 2011. [11] Buntine, W. L., & Weigend, A. S. (1991). Bayesian back-propagation. Complex Systems, 05(6), 603.

[12] D.J.C. MacKay, A practical Bayesian framework for backpropagation networks, Neural Comput. 4 (1992) 448}472.

[13] Bishop, C. (2006). Pattern Recognition and Machine Learning. Boston: Springer.

[14] MacKay, D. J. C. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems, 6 (3), 469-505.

[15] R.M. Neal, in: Bayesian learning for neural networks, Lecture Notes in Statistics, Vol. 118, Springer, New York, 1996.

[16] Pucheta, J., Patino, D. and Kuchen, B. "A Statistically Dependent Approach For The Monthly Rainfall Forecast from One Point Observations". In IFIP International Federation for Information Processing Volume 294, Computer and Computing Technologies in Agriculture II, Volume 2, eds. D. Li, Z. Chunjiang, (Boston: Springer), pp. 787–798. (2009).

[17] Pucheta, J., Patiño, H.D., Kuchen, B. (2007). Neural Networks-Based Time Series Prediction Using Long and Short Term Dependence in the Learning Process. In proc. of the 2007 International Symposium on Forecasting, New York, USA.

[18] Pucheta, J., Rodriguez Rivero, C., Herrera, M., Sauchelli, V. and J. Baumgartner . "Time Series Forecasting using Kernel and Feed-Forward Neural". XIV Reunión de Trabajo en Procesamiento de la Información y Control RPIC 2011, 16 al 18 de Noviembre de 2011 Oro Verde, Entre Ríos, Argentina. (2011).

[19] Dieker, T. (2004). Simulation of fractional Brownian motion. The Netherlands MSc theses, University of Twente Amsterdam.

[20] Abry, P.; P. Flandrin, M.S. Taqqu, D. Veitch. (2003), Self-similarity and long-range dependence through the wavelet lens. Theory and applications of long-range dependence, Birkhäuser, pp. 527-556.

Estimation of Water Quality Parameters Using the Regression Model with Fuzzy K-Means Clustering

Muntadher A. SHAREEF*, Abdelmalek TOUMI*
*Lab-STICC UMR CNRS 6285,ENSTA Bretagne
2, Rue François Verny, 29806 Brest Cedex 09, France

Ali KHENCHAF*
*Lab-STICC UMR CNRS 6285,ENSTA Bretagne
2, Rue François Verny, 29806 Brest Cedex 09, France

***Abstract*— the traditional methods in remote sensing used for monitoring and estimating pollutants are generally relied on the spectral response or scattering reflected from water. In this work, a new method has been proposed to find contaminants and determine the Water Quality Parameters (WQPs) based on theories of the texture analysis. Empirical statistical models have been developed to estimate and classify contaminants in the water. Gray Level Co-occurrence Matrix (GLCM) is used to estimate six texture parameters: contrast, correlation, energy, homogeneity, entropy and variance. These parameters are used to estimate the regression model with three WQPs. Finally, the fuzzy K-means clustering was used to generalize the water quality estimation on all segmented image. Using the in situ measurements and IKONOS data, the obtained results show that texture parameters and high resolution remote sensing able to monitor and predicate the distribution of WQPs in large rivers.**

Keywords*—In situ data measurements; IKONOS data; water quality parameters; GLCM; empirical models; fuzzy K-means clustering

I. INTRODUCTION

The use of remote sensing techniques to monitor, manage and predict water quality parameters (WQPs) and contaminants from the important things in recent years [1]. This as a result of the growing population, increasing industry, agriculture, and urbanization [2]. Where these techniques have helped to find and estimate pollutants in water with lower costs and greater potential [3]. It is known, that the contaminated water affects directly or indirectly on people's health, especially when it is a source or the only source of drinking water. Therefore, the water quality monitoring helps to assess quality of water bodies and identify contaminated areas [4]. In situ water quality measurement requires sampling which is expensive and time consuming in laboratory analysis.

For this reason, the remote sensing techniques can overcome these limitations by achieving an alternative means of water quality monitoring for larger average of temporal and spatial scales [5]. It should be noted that monitoring of water quality using remote sensing began in early 1970s depending on measure of spectral and thermal response in emitted radiation from water surfaces. In generally, empirical relationships between spectral properties and WQPs were established by the authors since in1974 and developed an empirical approach to estimate it [6]. The general forms of these empirical equations are:

$$Y = A + BX \; or \; Y = AB^{X} \qquad (1)$$

Where, Y is the remote sensing measurement vector (which includes: radiance, reflectance, energy …) and X is the water quality parameter vector of interest (i.e., suspended sediment, chlorophyll …), A and B are empirically derived factors [6].

Traditional methods for monitoring and estimating pollutants or water quality parameters WQPs [7] by optical satellite data were relied on the spectral response [8][9], thermal or scattering reflected from water [10][11], or by fusion spectral and microwave techniques [12]. Therefore, many references were recommended to use certain bands to find number of variables in waters. Thus, classification represents the nature of the separated, regardless of the classification accuracy. Therefore, in this study, we proposed different method to estimate WQPs using regression models on texture parameters. Thus, the proposed method, use the GLCM to estimate six texture parameters: contrast, correlation, energy, homogeneity, entropy and variance. Extracted texture parameters were corresponding to the ground-truth locations. Multiple regression models have been used to generate predictive models between texture parameters and WQPs. The predictive model with best correlation will be used later in the classification and identification of WQPs. Our work aims to study and estimate three parameters in water as WQPs: *PH* (is a measure of the acidity or basicity of an aqueous solution), phosphate (PO_4), and nitrate (NO_3) using an empirical equation as a function of extracted texture parameters. The empirical model, which has the highest accuracy, will be taken. Finally, to estimate the WQPs on the entire segmented water region in image, we apply the Fuzzy K-means classifier (FKM).

In the next sections, we present the study and the data used in this work. In section II, a simplified overview about study area has been introduced. Section III includes all methods used to get the purposes of this work. Then we will move on the results and discussion in section IV.

II. CASE STUDY

A. Study Area

The Tigris River is the eastern member of the two great rivers that define *Mesopotamia*, the other being the *Euphrates*. The river flows south from the mountains of southeastern Turkey through Iraq. The river Tigris is 1850 km in length, rising in the Taunus Mountains of Eastern Turkey. The total length of the river in Iraq is *1418 km* [13]. It consider main source for human use, especially for drinking water [14]. The study area represents the river Tigris within Baghdad city

(the capital of Iraq) and the length of river, extended 49 km from the Al-Muthana Bridge north Baghdad to the confluence with the Diyala river south Baghdad [15]. Fig. 1 level 1 illustrates the map study area of our work.

B. In Situ Data

In situ data measurements were collected from eight stations represents the main station of Baghdad city and distributed on Tigris river. This samples were collected from these stations in October 2012 and analysed in laboratories of ministry of environmental in Baghdad city to extract the water quality parameters included: PH , (PO_4) (NO_3) and other related variables. All parameters were done according to standard specifications presented by the American public health association [16].

C. IKONOS Data

Many types of satellites have the ability and potential appropriate for estimating WQPs. Higher resolution satellite is better in most cases, but signal-to-noise requirements of sensor technology impose limitations on the combined spectral, spatial and temporal resolutions for this reason no sensor can have a high spectral, high spatial and high temporal resolution. That mean if the pixel resolution of a sensor is small (high spatial resolution), the spectral bandwidth has to be large (low spectral resolution) to capture sufficient light energy for an acceptable signal-to-noise ratio. There is a trade off in spectral, spatial and temporal resolution and the best combination depends on the intended use of the sensor [17]. The IKONOS satellite was launched in September 24th, 1999 to provide global, accurate, high resolution imagery arrive to 1m [18]. In this study, one scene of IKONOS data was acquired on October 16th, 2012. The image was georeferenced to UTM, WGS48 and radiometrically corrected to minimize atmospheric effects. The image presented in Fig. 2 shows IKONOS image as input data in this work.

D. The Delineation and Extraction of River Water Image

The delineation and extraction of water bodies from remote sensing image is an important task useful for various applications such as, GIS database updating, flood prediction, and the evaluation of water resources [19]. Several techniques for the extraction of linear features from remotely sensed data have been introduced for high spatial resolution imagery [20]. The methodology and methods used to extract water area in satellite image can be summarized by three principal families of methods: Feature extraction method, supervised and unsupervised classification methods, feature based classifier and data fusion. Many researchers provided comprehensive overview on methods on water extraction (water segmentation) from high resolution satellite images. The authors in [21] provided comprehensive overview on methods on water extraction from high resolution satellite images. Fu June in [22] developed an automatic extraction of water body from TM image using decision tree algorithm which was adopted for the difference in spectral response from the water and terrestrial response.

III. The Methodology

The main methods of this study are expressed by Fig. 1.

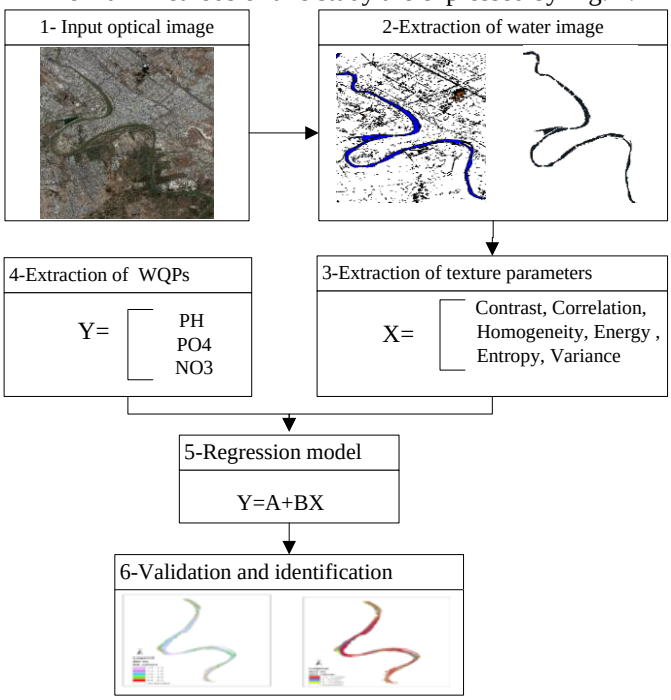


Fig. 1. Methodology adopted for WQPs estimation in this work

However, the difference in spectral response from the water and terrestrial response, the extraction and segmentation of water region in satellite images necessary for reasons: to ride of effect and to separate terrestrial area from the original image, easy to identifying, and advanced processing could be done easier and faster [23]. By using ENvironment for Visualizing Images (ENVI) and Geographical Information System (GIS), river has been extracted from the image after the segmentation step is applied on satellite image to extract three classes (land, vegetation and water) as shown in Fig 1 at level 2.

A. Texture Feature Extraction

There are many approaches used for texture analysis. We have chosen in this work, some parameters computed from the Gray-Level Co-occurrence Matrix (GLCM). The GLCM is the statistical approach for examining the textures that considers the spatial relationship of the pixels. The GLCM characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image [24]. It provides a second-order method for generating texture features to calculate the relationship between the conditional joint probabilities of all pairs of combinations of grey levels in the image parameters such as displacement d and orientation θ [25]. It can be calculated as symmetric or non-symmetric matrix. The symmetric of the GLCM is often defined a pair of grey levels (i, j) oriented at $\theta=0°$ and also be considered as being oriented at $\theta=180°$ [26].

Various texture features can be generated by applying GLCM statistics as in reference [24]. However, in our study six features (parameters) have been chosen and computed from the GLCM. These extracted parameters will be used to estimate the regression models to predict the QWPs. These are:

$$Contrast = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G} \sum_{j=1}^{G} P(i,j) \right\} \quad (2)$$

$$Correlation = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \quad (3)$$

$$Homogeneity = \sum_{i=1}^{G-1} \sum_{j=1}^{G-1} \frac{P(i,j)}{1+|i\_j|} \quad (4)$$

$$Energy = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{P(i,j)\}^2 \quad (5)$$

$$Entropy = -\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j) \times \log(P(i,j)) \quad (6)$$

$$Variance = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i-\mu)^2 P(i,j) \quad (7)$$

Where P is the matrix element, (i, j) intensities, G is the number of gray levels used, μ is the mean value of P . μ_x , μ_y , σ_x and σ_y are the means and standard deviations of the marginal-probability matrix [25].

To validate our model, we present in Table. 1 the texture parameters value computed from eight water regions (stations). The station (2, 4, 6 and 8) refers to training and used to estimate the models, while stations (1, 3, 5 and 7) refers to testing and used to validating our model. For each of these regions, the water is analysed and the corresponding WQPS *(PH, PO_4 and NO_3)* are extracted as shown in Tab II.

TABLE I. PARAMETERS EXTRACTED FROM IMAGE STATION

Station n°	contrast	correlation	energy	homogeneity	entropy	Data type
S2	0.0386	0.1203	0.9192	0.9807	0.2200	Training data
S4	0.1347	0.3215	0.6849	0.9326	0.6637	
S6	0.0715	0.2077	0.8433	0.9642	0.3367	
S8	0.0715	0.2077	0.8433	0.9642	0.0433	
S1	0.0396	0.1211	0.9170	0.9802	9.0511	Testing data
S3	0.0295	0.1495	0.9368	0.9853	9.0121	
S5	0.0715	0.2077	0.8433	0.9642	8.6706	
S7	0.2807	0.3914	0.3369	0.8596	7.1177	

B. Development of Multivariate Retrivel Algorithem

Most of remote sensing studies which are interesting in water quality parameters based on empirical models as we mentioned in the introduction. In this study multivariate algorithms using extracted texture parameters and satellite data have been done depend on equation (1) that is refer to multi-regression model. The statistical analysis depends on the WQPs and their corresponding texture parameters that shall be using in our work. In the multiple regressions, the independent variables were six texture parameters while dependent variables are water quality parameter to be calculated. All formulated model by empirical model essentially based on the correlation coefficient between measured data of water quality and texture parameters extract, regardless of whether the correlation is direct or indirect. When the correlation is high among the independent variables and WQPs, the predictive regression model will be strong, and the tendencies of values are high, as in Table II.

TABLE II. CORRELATION BETWEEN WQPS EXTRACTED AND PARAMETERS EXTRACTED FROM IMAGE STATION

	Contrast	Correlation	Energy	Homogeneity	Entropy	Variance
PH	0.845	0.876	-0.838	-0.846	0.354	-0.527
PO_4	-0.111	-0.076	0.117	0.110	-0.632	-0.933
NO_3	-0.751	-0.699	0.759	0.750	-0.993	-0.665

C. Validation of Multivariate Preductive Algorithms

1) Validation by comparison between measured and calculated WQPs

Measured WQPs refers to the observation were taken from the stations and calculated WQPs refers to the parameters calculated via satellite data. In order to obtain a strong validation, the validation applied for four different stations in first stage and in second stage all station was taken into account to find the difference in measured and calculated values.

2) Validation by fitting and confidence bounds models

Data fitting is the process of fitting models to data and analyzing the accuracy of the fit. Engineers and scientists use data fitting techniques, including mathematical equations and nonparametric methods, to model acquired data. The polynomial model has been selected to apply to analyzing and finding the errors. A polynomial is a function that can be written in the form:

$$P(x) = c_0 + c_1 x + \ldots\ldots + c_n x^n \quad (8)$$

For some coefficients $c_0, \ldots c_n$. If $c_n = 0$ then the polynomial is said to be of order *n*. A first order (linear) polynomial is just the equation of a straight line, while a second-order (quadratic) polynomial describes a parabola [26]. Confidence and prediction bounds define the lower and upper values of the associated interval, and define the width of the interval. The width of the interval indicates how uncertain you are about the fitted coefficients, the predicted observation, or the predicted fit. The confidence bounds for fitted coefficients are given by:

$$C = b \pm t\sqrt{S} \quad (9)$$

Where *b* are the coefficients produced by the fit, *t* depends on the confidence level, and is computed using the inverse of

Student's *t* cumulative distribution function, and *S* is a vector of the diagonal elements from the estimated covariance matrix

The simultaneous prediction bounds for the function and for all predictor values are given by:

$$P_{s,p} = y \pm f\sqrt{xSx^T} \quad (10)$$

Where *f* depends on the confidence level, and is computed using the inverse of the *F* cumulative distribution function. The Goodness of Fit (GOF) of a statistical model describes how well it fits into a set of observations. GOF indices summarize the discrepancy between the observed values and the values expected under a statistical model. To evaluate the goodness of fit, its required to calculate each of the Sum of Squares Error (SSE), R-square, adjusted R-square, and Root Mean Squared Error (RMSE).

D. Validation by Fuzzy K-Means Clustering

The Fuzzy K-means Clustering (FKM) algorithm performs iteratively the partition step and new cluster representative generation step until convergence. The applications of FKM can be founded in reference, which provided an excellent review of FKM. An iterative process with extensive computations is usually required to generate a set of cluster representatives [27]. Clustering a data set $X \subseteq R^N$ implies that the data set is partitioned into *k* clusters such that each cluster is compact and far from other clusters. One way to achieve this goal is through the minimization of the distances between the cluster center and the patterns that belong to the cluster. Using this principle, the hard k-means algorithm minimizes the following objective function [28]:

$$J = \sum_{k=1}^{K} \sum_{x_i \in F_k} d(m_k, x_i) \quad (11)$$

Where $d(m_k, x_i)$ is a distance measure between the center m_k of the cluster F_k and the pattern $x_i \in X$ Eq. (2) can be rewritten as

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_k(x_i) d(m_k, x_i) \quad (12)$$

Where $\mu_k(x_i) \in \{0,1\}$ is the characteristic function, i.e., $\mu_k(x_i) = 0$ if $x_i \notin F_k$, else $\mu_k(x_i) = 1$.When the clusters are overlapping, each pattern may belong to more than one cluster, i.e., $\mu_k(x_i) \in [0,1]$. Hence, $\mu_k(x_i)$should be interpreted as a membership function rather than the characteristic function. Therefore, the objective function (3) can be modified to the following:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_k^q(x_i) d(m_k, x_i) \quad (13)$$

Where $\mu_k(x_i) \in [0,1]$a fuzzy membership function and q is now is a constant known as the index of fuzziness that controls the amount of fuzziness.

IV. Result and Discussion

A. Analysis of Correlation

Scattering pattern has been studied in the early stages of the work, for two main reasons: to find the correlation between WQPs and extracted texture parameters. The forms of scattering indicate the relationship between the parameters in direct and indirect. It is not important what kind of relationship and behavior was done, because the correlation takes the absolute value to determine the strength between parameters. Examination of the correlations between the parameters that extracted by method of texture analysis and measured from the station shows in Table II. Where, there is a strong direct correlation between *PH* and two texture parameters: contrast, correlation (0,845, 0,876) respectively. That means, these two parameters influenced more directly with *PH* . If the *PH* increase the two texture parameters will increase and vice versa. In the same time, It was indirect correlation between *PH* and energy and homogeneity (-0,838, -0,846) respectively. This interprets inverse relationship will increase with decrease. En general in both cases there is strong correlation. A weak correlation was found between *PH* , entropy and variance.The high correlation between extracted parameters probably means that these texture parameters are measuring similar aquatic properties. As for the second parameters (PO_4), which is one of the important pollutants in the water. PO_4 was found a high correlation with variance. The correlation does not appear with other studied texture parameters. The NO_3 , which represents the purity in the water, showed a high correlation with entropy.

B. Generation of Multivariate Predictive Algorithms

Using multiple regression model making possible to predict eight equations to measure *PH* according to the type of texture used with average of accuracy (95%).

$$PH_{calculated} = 7.167 + 5.791 \times C \quad (14)$$

$$PH_{calculated} = 6.998 + 2.924 \times Co \quad (15)$$

$$PH_{calculated} = 6.750 - 10.103 \times C + 7.807 \times Co \quad (16)$$

$$PH_{calculated} = 9.144 \times E \quad (17)$$

$$PH_{calculated} = 22.748 \times E - 15.345 \times E^2 \quad (18)$$

$$PH_{calculated} = 7.93 \times H \quad (19)$$

$$PH_{calculated} = 27.748 \times H - 20.619 \times H^2 \quad (20)$$

$$PH_{calculated} = -4.792 \times E + 12.044 \times H \quad (21)$$

Where contrast (C), correlation (Co), energy (E), homogeneity (H). Each predicted equation has an accuracy corresponding to R^2 and probability value in regression analysis model. Hence, the equations from 1-8 have (0.8450, 0.8760,0.895, 0.9830, 0.9998, 0.9970, 0.9990, 0.9997) respectively. Equation (17) and (19) have been excluded accuse of probability values were higher than 0.05. Equation (20) have been chosen to represent the classification because of high accuracy.

Using equation (20) does not prevent using other equations according to the accuracy and type of the texture available. For PO_4 and NO_3 there was strong relationship with variance V and entropy (En) respectively expressed by:

$$PO_4 = 4.927 - 0.316 \times V \quad (22)$$

$$NO_3 = 14.344 - 16.837 \times En \quad (23)$$

C. Validating of Predictive Algorithems

As it mentioned above, three type of validation has been done to measure the strength of equations. Analysis confidence bounds models have been done to measure the accuracy of all productive algorithms as shown in Figs.2 and 3. Polynomial quadric model has been demonstrated to fit measured and calculated PH , PO_4 and NO_3 .The result shows that all points fall into boundary of confidence equal to (95%).The indications of goodness of fit (GOF) has been also calculated as; SSE= 0.02149, R-square= 0.9853, adjusted R-square=0.9486 and RMSE= 0.1037and all of these indicators give high quality.

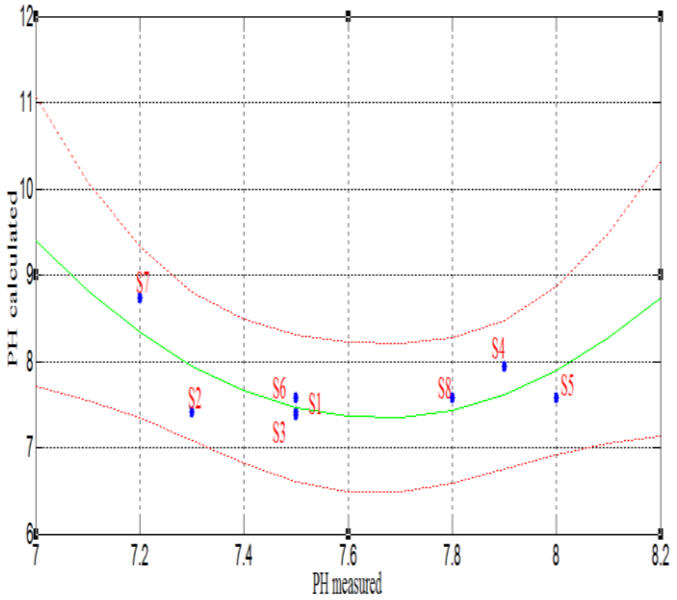


Fig. 2. Fitting and confidence bounds models of PH.

D. Image Segmentation and Validation

Image segmentation using fuzzy K-means result shows in Figs.6 and 7. The fuzzy logic is classified the mixed pixel to specific category based on the descriptions of the input and output variables. Fuzzy logic rules applied to incorporate expert knowledge. Fixing a set of rules has been done to classify PH image. Three classes have been selected to represent PH .

Figs.4and 5, show the result of segmentation and distribution of PH and NO_3 . All of these parameters were full into under safety factors and corresponding to ground measurements.

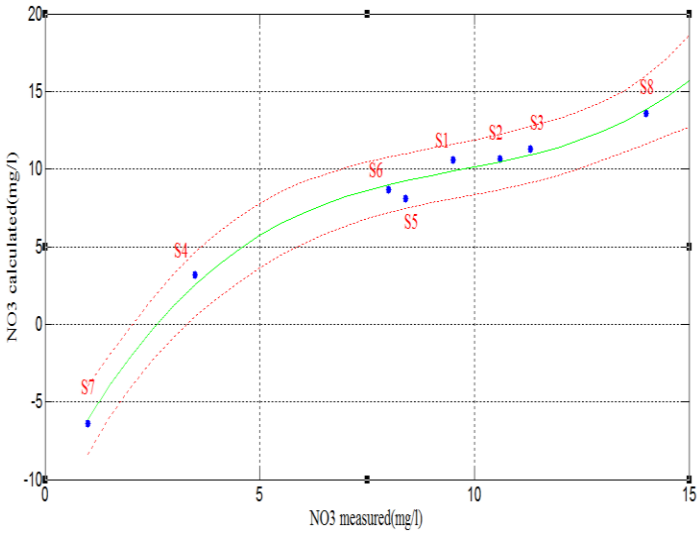


Fig. 3. Fitting and confidence bounds models of NO_3 .

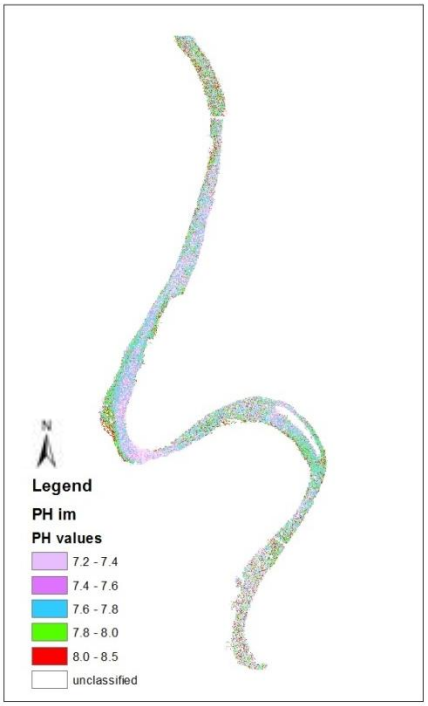


Fig. 4. PH distrubution map

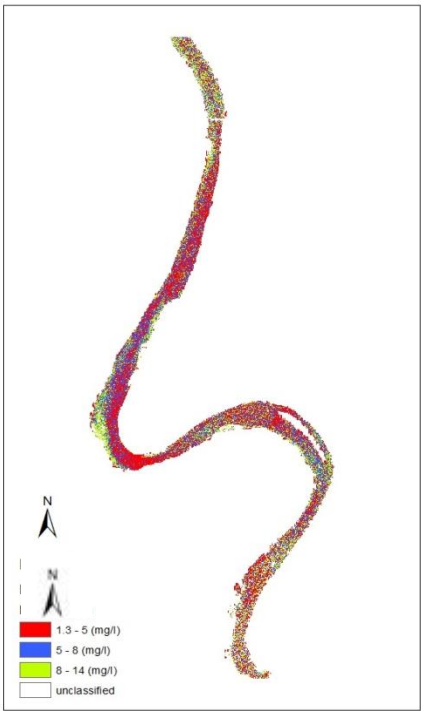


Fig. 5. NO_3 distrubution map

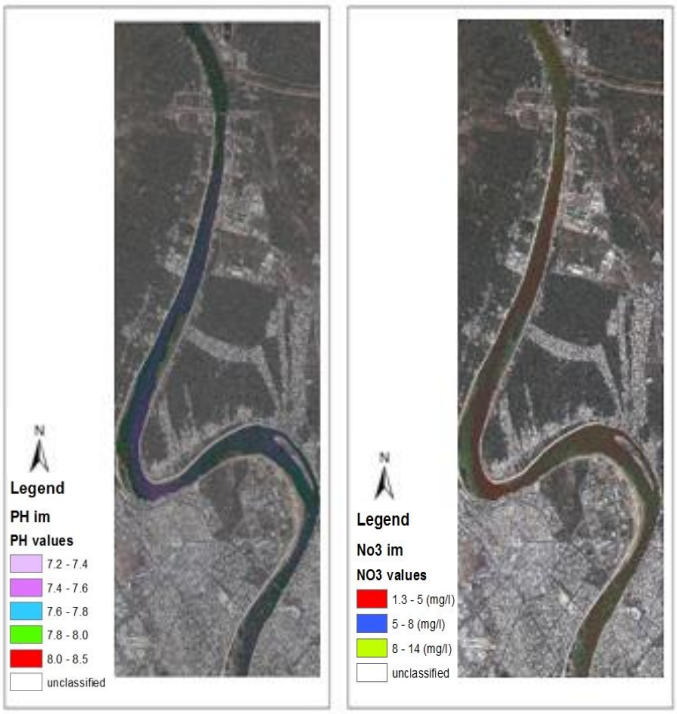


Fig. 6. Concentration of PH and NO_3 .

V. Conclusion

In this study, potential applications for assessing and monitoring water quality, using texture parameters have been demonstrated using GLCM. Texture parameters were extracted for each sample corresponding to the ground-truth locations. Homogeneity, entropy and variance were found to be the most suitable texture parameters for predicting PH , PO_4 , and NO_3 concentrations using empirical models with high correlation. This method helped to calculate PH from many equations according to texture parameters and with different accuracies. Confidence bounds models have indicated the substantial convergence between measured and calculated variables. Some of the points possessed very high values of pollution which caused a large gap in the homogeneity values of the points. This gap or disparity in the measured values did not affect the accuracy of the model. Analysis between remotely sensed data and ground data have indicated the possibility to mapping two of WQPs, expect the third water parameter PO_4 which had zero in image texture. For this reason, it is ignored from results. Using fuzzy K-mean method was helped the rules about the texture input and description of classes to get good classification for studied parameters.

As a perspective work, the future research should contemplate this issue by selecting more number of sampling stations in proper locations so that more accurate results can be obtained. As well as this method could be applying with different type of satellite images and compared it with other methods especially which are concern to study the roughness of the surfaces and backscattering models.

Acknowledgment

The authors would like to thank the Iraqi Ministry of Environment to provide all the facilities to get the information about our work. They also thank the GIS Center in the ministry and especially the responsible of the informatics department Mss. Rua Khalid for their helpful and their continued cooperation. Also, the authors would like to thank Campus France for their support in making this work.

References

[1] N. Usali,and M. Hasmadi, “Use of remote sensing andGIS in monitoring water quality,”Journal of sustainable sevelopment, Selangor,vol. 3, no. 3, September 2010.

[2] W. He, S. Chen, X. Liu,and J. Chen, “Water quality monitoring in slightly-polluted inland water body through remote sensing A case study in Guanting reservoir,Beijing, China,”Higher education press and Springer-Verlag, China, vol. 2, pp. 163-171, Beijing 2008.

[3] F. L. Hellwegera, P. Schlossera, U. Lalla, andJ.K. Weisselc, “Use of satellite imagery for water quality studies in New York Harbor,”estuarine, coastal and shelf science,vol. 61, pp. 437–448, June2004.

[4] A. M. Sheela, J. Letha, S. Joseph, and K. K. Ramachandran, “Prediction of water quality of a lake system by relating secchi disk depth and IRS-P6 radiancedata,” International conference on technological trends., Trivandrum, November 2010.

[5] A. Mumtaz Bhatti, “Modelling and monitoring of suspended matter in surface waters using remotely sensed data,” Thesis, Kochi University of Technology, Japan, March 2008.

[6] C. Jerry, V. Zimba, and H. Everitt,“ Remote sensing techniques to assess water quality,” Journal of Photogrammetric engineering and remote sensing,vol. 69, no. 6, pp. 695-704, June 2003.

[7] F. Karimipour, M. R. Delavarand M. Kinaie, “Water quality management using GIS data mining,” Journal of Environmental informatics, vol. 5, pp. 61-72, 2005.

[8] C. Giardino, V. E. Brando, A. Dekker, N. Strömbeck, and G.Candiani,“Assessment of water quality in Lake Garda (Italy) using Hyperion,” Journal of remote sensing of environment, vol. 109, pp. 183–195, December 2006.

[9] E. Brando, and G. Dekker, "Satellite hyperspectral remote rensing for estimating estuarine and coastal water quality," IEEE transactions on geoscience and remote Sensing, vol. 41, pp. 1378-1386, June 2003.

[10] Y. Zhangh,J. Pulliainen, S. Koponen, and M. Halliainen,"Empirical algorithm of secchi disk depth using optical and microwaves remote sensing datafrom the gulf of finland and the Archipelago sea," Boreal enviromental research, vol. 8,pp. 251-261, Helsinki, September 2003.

[11] K. Cheng, and T.Chiang Lei, "Reservoir trophic state evaluation using Landsat TM images," Journal of the american water resources,vol. 37,pp. 1321-1334, 2001.

[12] Y. Zhang, T. Pulliainen, S. Koponen, andT. Hallikainen,"Water Quality Retrievals From CombinedLandsat TM data and ERS-2 SAR data in the gulf of Finland," IEEE Transactions on Geosciences and Remote sensing, vol. 41,pp. 622-629. March 2003.

[13] M.V. Mikhailova, "The hydrology, evolution, and hydrological regime of the mouth area of the Shatt al-Arab River". Water Resources 36 (4): 380-395.2009.

[14] A. Rabee,and A. Ahmed, " Seasonal variations of some ecological parameters in tigris river water at Baghdad region, Iraq," Journal of water resource and protection, vol. 3, pp. 262-267, April2011.

[15] A. Ali, N. Al-Ansari, and S. Knutsson, "Morphology of Tigris river within Baghdad city," Journal of hydrology and earth systym science,vol. 16, pp. 3783–3790, September 2012.

[16] American Public Health Associ, "Standard methods for the examination of water and wastewater," 20th edition, Washington, DC, 1998.

[17] Y.Fong,J. Liou, J. Hou , W. Hung , S. Hsu, Y. Lien , M. Daw,K. Sheng and Y. Wang, "A multivariate model for coastal water quality mapping using satellite remote sensing images," Sensors, vol. 8, pp. 6321-6339, October2008.

[18] G. Dial, H. Bowen, F. Gerlach, J. Grodecki, and R. Oleszczuk, "IKONOS satellite, imagery, and products," Journal ofremote sensing of environment, vol. 88, pp. 23–36, August2003.

[19] V. Shah, A. Choudhary, and K. Tewari," River extraction from satellite image ," International journal of computer science issues, vol. 8, no. 2, July 2011.

[20] N. Dinh duong, "Water body extraction from multi spectral image by spectral pattern analysis,"Journal ofPhotogrammetry, remote sensing and spatial information sciences, Melbourne, vol. XXXIX-B8, September 2012.

[21] R. Kumar, and S. Deb, "Water-body area extraction form high resolution satellite images – An introduction, review,and comparison," International journal of image processing, vol.3,pp. 353-372, September 2010.

[22] F. June, W. Jizhou, and L.Jiren, "Study on the automatic extraction of water body from TM image using decision tree algorithm," Proc. of SPIE,vol. 6625, pp. 662502-1:662502-5, 2008.

[23] V. Amandeep, "Identification of landand water regions in a satellite image: a texture based approach," IJCSET, vol. 1, pp. 361-365, August 2011.

[24] N. Zulpe, and V. Pawar, "Glcm textural features for brain tumor classification," IJCSI International journal of computer Science Issues, vol. 9, no. 3,pp. 354-359, May 2012.

[25] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification,"IEEE Transaction on system, vol. 6, pp.610–621, 1973.

[26] M. Arebey, M. A. Hannan, H. Basri,and R. A. Begum, "Bin level detection using gray level co-occurrence matrix in solid waste collection,"Proceedings of the world congress on engineering and computer science, vol. 2, San Francisco, USA, October 2012.

[27] C. Tang Chang,Z. C. Jim,and M. Jeng, "A fuzzy k-means clustering algorithm using cluster center displacement," Journal of information science and engineering, vol. 27, pp. 995-1009, 2011.

[28] M.Sarkar, and T. Yun, "Fuzzy k-means clustering with missing values," AMIA Journal of the American Medical Informatics Association, pp. 588-592, 2001.

A Compound Generic Quantitative Framework for Measuring Digital Divide

Noureldien A. Noureldien
Department of Computer Science
University of Science and Technology,
Omdurman, Sudan

***Abstract*—The term digital divide had been used in the literature to conceptualize the gap in using and utilizing information and communication technologies. Digital divide can be identified on different levels such as individuals, groups, societies, organizations and countries. On the other hand, the concept of e-Inclusion is coined to define activities needed to bridge digital divide.**

One of the most challenging research areas in digital divide that had been a subject for exhaustive studies is measuring digital divide. Researchers have proposed many metrics and indices to measure digital divide. However, most of the proposed measures are bivariate comparisons that reduce measurement to comparisons of Internet penetration rates or alike.

This paper proposes a compound generic framework for quantitative measuring of digital divide on the individuals or group level. The proposed framework takes into account the context of the digital divide in each society.

Keywords*—*Digital Divide; Digital Divide Indicator; E-inclusion; Inclusion Factors; Inclusion Activities

I. Introduction

The term Digital Divide is coined in 1995 and popularized in the late 1990s to describe the social division among people in terms of their involvements of using information and communication technologies [1].

A widely accepted definition of digital divide is the one provided by the Organization for Economic Co-operation and Development (OECD): ‘‘the term digital divide refers to the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard both to their opportunities to access ICT and to their use of the Internet for a wide variety of activities.

The digital divide may appear due to historical, socioeconomic, geographic, educational, behavioral, or generation factors, or due to the physical incapability of individuals [2].

The underlying reasons for the increasing attention to the term are lies in the wide recognition that ICT and specifically the Internet have both empowering and discriminating ability to the citizens of a society [3].

Although the digital divide was initially understood in a binary way, that is to say, a choice between ‘‘has’’ and ‘‘has not’’ access to ICT, which is very reductive, imprecise, and inaccurate, today digital divide is understood to be a complex, multidimensional phenomenon [4][5].The digital divide, therefore, represents “a major challenge” confronted by any information-based network society and knowledge economy.

In analyzing the digital divide phenomena, researchers took different levels of analysis. Three different levels of digital divide analysis are frequently explored, namely, global level, organizational level and individual level [6]. And in terms of the types of inequality that results from the digital divide there are at least two inter-related but conceptually different types of divide, ICT access inequality and ICT use inequality.

During last decade many voices were raised calling for the necessity of defining quantification of digital divide, that is to say, a reliable measurement and analysis of the digital divide is desperately needed.

This paper proposes a compound generic framework for quantitative measuring of digital divide based on a stated theory that correlate digital divide and e-inclusion concepts. While the author believes on the appropriateness of this approach in measuring different levels of analysis, the paper highlights only the individual or group level analysis.

The rest of this paper is organized as follows; in section 2 related research works was discussed. In section 3 the theoretical concepts and notation related to the developed framework was stated. Section 4 presents the proposed framework. Finally conclusions are given in section 5.

II. Related Work

The research on measuring digital divide is basically either a demonstration of the existence and magnitude of digital divide, or it focuses on the identification of various determinants of such divides. Some researchers apply statistics to quantify digital divide. In [7] the authors employ the ratio of deviation to mean as the indicator of divide magnitude. In [8] authors also took the same criterion to assess the level of national digital divide, while in [9] authors employ this approach to evaluate the magnitude of global digital divide.

It had been argued that the Gini coefficient should be especially suitable as a standard measure of digital divide [10], and it is used as standard measure to quantify the magnitude of digital divide on individual level [11].

Most of these approaches measure the difference in Internet Usage between individuals or social groups by using either a bivariate analysis (such as Age less than 40 and age >40, Urban and rural …etc) and use that difference as a measure for the digital divide [12][[5][13], or they use log linear modeling in order to simultaneously analyze a number of variables, called multivariate approach (such as measuring the difference in Internet usage between those who are educated and less than 16 years in Urban and rural areas, and take that difference as a measure for digital divide [14][9][15].

To overcome the problem resulted from the fact that static indicators are not sensitive to changes in the corresponding absolute magnitude of the indicator growth rates [16], the distance time methodologies are proposed as a new statistical measure in dynamic gap analysis [17]. In this new approach the levels of variable(s) are used as identifiers and time is the focus of comparison.

In [18] authors proposes three essential approaches to measure digital divide; they use loglinear modeling to address the interactions among the factors affecting the digital divide. Second, they use compound measures that integrate a number of variables into a single indicator. Third, they apply time-distance methodology to analyze changes in the digital divide.

Sometimes digital divide indicators take different or even contradictory values, to justify these indicators a compound measure can be used. Recently, various compound ICT measures have appeared, such as the Technology Achievement Index [19], the Information Society Index [20], the Internet Connectedness Index [21], Digital Access Index [3], the Networked Readiness Index [22], and the Digital Opportunity Index [23]**.**

One of the more sophisticated examples of compound measures is the Digital Divide Index (DIDIX) which was developed within the Statistical Indicators Benchmarking Information Society (SIBIS) project, an EU research framework program led by Empirica [24].

Authors in [25] show a digital divide index DDIX, in which they have compared the technology adoption among risk groups to the adoption among the population average as a measure for the digital gap.

The proposed framework is a compound digital divide index for individual level. It is a general framework that can be applied to measure digital divide in any societal context, and it uses the e-inclusion theory to formulate a quantitative measure for digital divide.

III. A Compound Generic Quantitative Framework

This paper deals with digital divide between individuals or social groups, thus its unit of observation is individuals or social groups, and since the approach is a generic one no specific independent variables (such as age, gender, education, income, …) or specific digital divide indicators (such as Infrastructure, Access, e-skills, Internet Usage, ….), are specified.

Only some indicators are used as examples for illustrative purposes.

The proposed measure or index will take into account the context of the digital divide in each society, that is; each society will have its own indicators and sub indicators which have weights that reflects the society context. For example, in developing countries societies, infrastructure may be defined as an indicator with high weight, while in developed countries societies that have already good infrastructure, infrastructure may not be defined as an indicator or may have a very low weight. To define the proposed quantitative framework, the following definitions and notations are introduced.

Digital Divide Indicator

A digital divide indicator defines a gap that prohibits an individual or a social group from active participation in the e-Society and can be used as a measure for defining digital exclusion. Examples of digital divide indicators are: Infrastructure, Access, Internet Usage and E-skills. The set of digital divide indicators for a given society group will be denoted by $Q = \{q_1, q_2, q_3, \ldots q_n\}$.

Inclusion factors

These are the societal, economical and technical factors that mitigate or eliminate the exclusion caused by a specific indicator. For example Access indicator might has the following inclusion factors: Availability of broadband, Availability of access devices (desktop, laptop, pad, phone …etc), Affordability of Internet access prices and Basic ICT skills (editing, email, web browsing, search engines)

The set of inclusion factors for an indicator qi will be denoted by $Y_{qi} = \{y_1, y_2, \ldots . y_t\}$.

Inclusion Activities

These are activities initiated by public, private sector and civil society to provide individuals and societies with a specific inclusion factor. For example for the Access inclusion factor "Affordability of Internet access prices", the inclusion activities may include:

- providing access motivations,
- providing employment opportunities and
- providing low access prices.

The set of inclusion activities corresponding to a single inclusion factor y_j will be denoted by $s_j = \{x_1, x_2, \ldots x_k\}$ where x_m ,$1 <= m <= k$, is an inclusion activity.

Consequently the set of all inclusion activities corresponding to a digital divide indicator q_i was denoted by S_{qi} where $S_{qi} = \{s_1, s_2, \ldots s_t\}$.

Absolute Inclusion Factor Weight

Inclusion factors associated with a specific indicator may have different strength and influence in mitigating digital exclusion. This strength and influence is referred to as the inclusion factor weight. The weight given to each factor should reflects the society context and should be a signed by experts. Also the assigned weight value for a given inclusion factor should consider the cost and time of the inclusion activities corresponding to that factor.

One possible definition for such weight may be given by

$|yi| = (\sum_{i=1}^{k} (\text{Cost}(xi)))/k$…… (1)

where Cost (xi) is a numerical adjusted value (0-100 for example) that mapped the financial cost and time needed to provide xi.

This inclusion factor weight will de referred to as the absolute inclusion factor weight.

Consequently, the total sum of inclusion factors absolute weights for an indicator q_j is given by

$|q_j| = \sum_{i=1}^{t} |yi|$…….. (2)

This total sum of weights assigned to inclusion factors must may be accumulated to 1 or 100.

For example, for Access indicator the absolute inclusion factors weights might be: 40 for availability of access devices, 30 for affordability of Internet access prices and 30 for basic ICT skills.

Inclusion factor gained weight

An individual or a social group may lose or gain partially or fully an inclusion factor. For example for the inclusion factor "Affordability of Internet Access Prices" an individual may be living in urban area where broadband is available but he has a job with low income that does not allow him to afford Internet access prices.

This partially or fully losing or gaining was defined as *an inclusion factor gained weight*. A complete lose of an inclusion factor is evaluated to 0 weight, full gaining is evaluated to the absolute factor weight of the inclusion factor, while the weight of partially gained should be evaluated in correspondence to the inclusion activities needed to achieve the absolute weight, using the Cost function for example.

For example an individual gained weights of inclusion factors of Access Indicator might be: 40/40 for availability of access devices, 20/30 for affordability of Internet access prices and 10/30 for the basic technical ICT skills.

The gained weight values of the inclusion factor y_i is denoted by $|y'_i|$. The difference (d_i), between the absolute inclusion factor weight value $|y_i|$, and the corresponding gained value $|y'_i|$, is denoted by:

$$d_i = | y_i - y'_i| \quad \text{…….. (3)}$$

defines the gap weight of the inclusion factor y_i.

Consequently, the total weight of gained weigh values by an individual /social group for an Indicator q_j is given by

$|q'_j| = \sum_{i=1}^{t} |y'i|$ …….. (4)

Now, from equations (2) and (4)

$$|q_j - q'_j| = \sum_{i=1}^{t} |yi - y'i| \quad \text{……..……. (5)}$$

which defines the total gap weight for a specific indicator q_j. Thus the total digital divide weight |DD| is given by

$|DD| = (\sum_{j=1}^{\eta} | qj - q'j |) /n$ ……..……..…. (6)

Equation (6) gives a quantitative measure for the digital divide (DD) assuming that all indicators contribute equally to the digital divide.

If indicators contribute with different weights to DD, for example if a digital divide DD is measured using the indicators; Infrastructure, Access, Internet Usage and E-skills, then the contribution of these indicators to the value of DD may be 40%, 30%, 15% and 15% respectively. The percentage of contribution of indicator q_j is denoted by αj%. Consequently

$|DD| = (\sum_{j=1}^{\eta} ((| qj - q'j | * \alpha j)/100))$ ……..… (7)

IV. Discussion

The above theory shows a generic framework for quantitative measuring of digital divides. The framework relays on three levels; indicators, inclusion factors and inclusion activities.

The cost of inclusion activities plays the major role in determining the weight of its associated inclusion factor, and consequently inclusion factors weight determine the weigh their corresponding indicator contribute to whole digital divide.

V. Conclusion

Measuring digital divide is a challenging problem. Quantitative measures always have a sounding essence. This paper contributes to the theory of digital divide and proposes a generic framework for quantitative measuring of digital divide. The internal structure of the framework entails flexibility that allows considering the context of digital divide of any society and proofs its applicability in all digital divide analysis levels; global level, organizational level and individual level.

References

[1] Hawkins, S.. Beyond the digital divide: Issues of access and economics. *The Canadian Journal of Information and Library Science, 29 (2),* 171-189.2005.

[2] Cullen, R. Addressing the digital divide. *Online Information Review 5*:311–320. 2001.

[3] Guillen, M. F. & Suarez, S. L.. Explaining the global digital divide: Economic, political and sociological drivers of cross-national Internet use. *Social Forces, 84(2),* 681-708. 2005.

[4] Bertot, J.C., The multiple dimensions of the digital divide: more than technology 'haves' and 'haves nots', *Government Information Quarterly 20, 2003,* pp.185–191, 2003.

[5] Hsieh, A., Rai, A., & Keil, MUnderstanding digital inequality: comparing continued use behavioral models of the social-economically advantaged and disadvantaged, MIS Quarterly 32, 2008, pp. 97–126.

[6] Dewan,S., Riggins, F. J.The digital divide:Current and future research directions. *Journal of the Association for Information Systems, 6(12),* 298-336. 2005.

[7] Cole, J. I., Suman, M., Schramm, P., Lunn, R., Aquino, J.–S., and Lebo, H. The digital future report: Surveying the digital future,year four. Ten years, ten trends. Retrieved from http://www.digitalcenter.org/downloads/DigitalFutureReport-Year4-2004.pdf

[8] Jin, J. & Xiong, C. Digital divide in terms of National Information Quotient: The perspective of Mainland China. Paper presented on International Conference on The Digital Divide: Technology and Politics in the Information Age, 22-23 August 2002, Hong Kong.

[9] Corrocher, N., Ordanini, A. Measuring the digital divide: A framework for the analysis of cross country differences. *Journal of InformationTechnology,17*,9-19. 2002.

[10] Chakraborty, J. &Bosman, M. M. Measuring the digital divide in the United States: Race, income,and personal computer ownership.*The Professional Geographer, 57 (3)*, 395-410. 2005.

[11] Jianbin Jin & Angus Weng Hin Cheong. Measuring Digital Divide: The Exploration in Macao, *Observatorio (OBS*) Journal, 6 (2008)*, 259-272. 2008.

[12] Bell, P., Reddy, P. and Rainie, L. Rural Areas and the Internet. Retrieved from. http://www.pewinternet.org/pdfs/PIPRural Report.pdf,

[13] Kalkun, M., and Kalvet, T. Digital divide in Estonia and how to bridge it. Tallinn: Emor and PRAXIS Center for Policy Studies. 2002.

[14] Cava-Ferreruela, I., Alabau-Munoz, A. Key constraints and drivers for broadband development: A cross-national empirical analysis. Presented at the 15th European Regional Conference of the International Telecommunications Society (ITS), Berlin, Germany, September.2004

[15] Grigorovici, D. M., Constantin, C., Jayakar, K., Taylor, R. D., and Schement, J. R. InfoMetrics: A structural equation modeling approach to information indicators and "e-readiness" measurement. Paper presented at the 15th Biennial Conference of the International Telecommunication Society (ITS), Berlin. 2004.

[16] Sicherl, P. Different statistical measures provide different perspectives on digital divide. Paper presented at the 6th Conference of the European Sociological Association, Murcia.2003.

[17] Sicherl, P. A new generic statistical measure in dynamic gap analysis. The European e-Business Report. Luxembourg: European Commission. 2004.

[18] VasjaVehovar et al, Methodological Challenges of Digital Divide Measurements. *The Information Society, 22:* 279–290. 2006.

[19] United Nations Development Program. (2001). Human development report 2001.New York: Oxford University Press.

[20] IDC. (2001). The IDC/World Times Information Society Index: The future of the information society. Framingham, MA: IDC

[21] Jung, J.-Y., Qiu, J. L., and Kim, Y.-C. Internet connectedness and inequality: Beyond the "divide." *Communication Research* 28(4):507–535. 2001.

[22] Dutta, S., and Jain, A. *The* Networked Readiness Index 2003–2004: Overview and analysis framework, 2004.

[23] International Telecommunication Union. Measuring digital opportunity. Paper presented at the WSIS Thematic Meeting on Multi-Stakeholder Partnerships for Bridging the Digital Divide, Seoul, Republic of Korea, June. 2005.

[24] Empirica, Communication and Technology Research. (2005). Retrieved from http://www.empirica.biz

[25] Tobias, H and Hannes, S. The Digital Divide Index – A measure of Social Inequalities in the Adoption of ICT, ECIS June 6–8,2002 Gdańsk, Poland.

XCS with an internal action table for non-Markov environments

Tomohiro Hayashida
Graduate School of Engineering,
Hiroshima University,
1-4-1, Kagamiyama, Higashi-Hiroshima,
Hiroshima, 739-8527, JAPAN
Email: hayashida@hiroshima-u.ac.jp

Ichiro Nishizaki
Graduate School of Engineering,
Hiroshima University,
1-4-1, Kagamiyama, Higashi-Hiroshima,
Hiroshima, 739-8527, JAPAN
Email: nisizaki@hiroshima-u.ac.jp

Keita Moriwake
Graduate School of Engineering,
Hiroshima University,
1-4-1, Kagamiyama, Higashi-Hiroshima,
Hiroshima, 739-8527, JAPAN

***Abstract*—To cope with sequential decision problems in non-Markov environments, learning classifier systems using the internal register have been proposed. Since, by utilizing the action part of classifiers, these systems control the internal register in the same way as choosing actions to the environment, they do not always work well. In this paper, we develop an effective learning classifier system with two different rule sets for internal and external actions. The first one is used for determining internal actions, that is, rules for controlling the internal register. It provides stable performance by separating control of the internal register from the action part of classifiers, and it is represented by "If [*external state*] & [*internal state*] then [*internal action*]," and we call a set of the first rules the internal action table. The second one is for selecting external actions as in the classical classifier system, but its structure is slightly different with the classical one; it is represented by "If [*external state*] & [*internal state*] & [*internal action*] then [*external action*]." In the proposed system, aliased states in the environment are identified by observing payoffs of a classifier and referring to the internal action table. To demonstrate the efficiency and effectiveness of the proposed system, we apply it to woods environments which are used in the related works, and compare the performance of it to those of the existing classifier systems.**

***Keywords*—Learning classifier systems; Non-Markov environments; XCS; Internal register.**

I. Introduction

Although classifier systems with if-then rules which develop through interaction with environments were initially considered as a computational model for cognition [12], [14], they are now widely applied to many areas, including autonomous robotics [8], [29], classification and data mining [33], [25], [15], traffic signal control [2], [4], and FPGA design [6].

A framework of classifier systems was initially proposed by Holland [11], [12], and subsequently a wide variety of classifier systems have been developed [7], [31], [32]. Especially, XCS developed by Wilson [32] has been attracting a lot of attention, and it is publicly recognized as one of the most successful learning classifier systems. Before XCS, the fitness of a classifier was calculated by using the expected payoff or the strength in the traditional learning classifier systems, and therefore there was a problem that classifiers which have low expected payoffs but are required to find optimal policies are eliminated by the procedure of genetic algorithms. To overcome this difficulty, the degree of accuracy is used as the fitness in XCS, and it is based on the difference between the predicted payoff and the actually received payoff.

In this paper we deal with non-Markov environments or partially observable Markov decision processes. In Markov environments where the probability of being in a given state depends on the current state and action but not on any past states or actions, agents can select the optimal policy by appropriately utilizing the information of the environment. If even in a Markov environment an agent can obtain only restrictive information of the environment, such a process is called a partially observable Markov decision process (POMDP). In a POMDP, different states can exist even if agents obtain the same information from the environment, and then the agents are said to suffer from a perceptual aliasing problem. In an aliased position or state, an agent cannot identify the current situation only through the information obtained from the environment by itself, and then it cannot select the next optimal action. From this reason, one can understand that the learning method of an agent in non-Markov environments is similar to that of POMDPs.

Since XCS determines an action by using the information about the environment at the current period, it is difficult to select an appropriate action in a non-Markov environment involving aliased states which cannot be discriminated only by the information about the environment at the current period. Several attempts using reinforcement learning and learning classifier systems for finding optimal policies in a non-Markov environment or a POMDP have been reported. For instance, Pineau et al. [22] propose an algorithm based on reinforcement learning for POMDPs, and apply it to a robot domain problem where an agent searches for and tags a moving opponent. Roy et al. [24] try to solve large scale POMDPs problems by reducing the dimensionality of the problem space. Shani et al. [26] present a learning model for POMDP based on reinforcement learning with memories of tree structure. Methods based on classifier systems such as ZCS [5] and ACS [27] have been also developed and applied to the grid-like woods environments which are benchmark problems for POMDPs. Moreover, Lanzi and Wilson [20] develop XCSM and XCSMH which are extensions of XCS, and intends to resolve environmental aliasing by incorporating the internal registers. In XCSM, both an external action which means an action that the agent takes in the environment and an internal action for controlling the internal register are specified in the

action part of a classifier, and they are treated in the same way. Although implementation for using the internal register is simple and elegant, its performance is not always good as we will show the experimental result. Moreover, it is difficult to determine an appropriate size of the internal register, and if it is too large for a given problem, the space of exploration becomes larger than necessary. Recently, Hamzeh et al. [10] develop the parallel specialized XCS (PSXCS), Zang et al. [35] develop XCS with average reward (XCSAR) which the Q-learning employed by XCS is replaced to R-learning not to limit the length of action chains. Preen and Bull [23] introduce discrete and fuzzy dynamical system within XCSF learning classifier system [34].

In PSXCS, along the lines of the history window approach [16] the information of the environments and the selected actions are recorded and aliases states are identified by the condition part of classifiers corresponding to the history of the environments and the selected actions.

Reinforcement learning is a type of machine learning such that an agent selects an action in an environment so as to maximize the cumulated sum of reward function. The agent receives the reward from the environment after taking an action, and by repeating this procedure it learns to take an appropriate policy so as to maximize the reward. In non-Markov environments or POMDPs, the agent cannot always obtain the optimal policy through the usual implementation of reinforcement learning. By using some ideas such as referring to the history of actions which are taken by the agent and the perceived information about the environment or reducing the dimensionality of the problem space, systems of reinforcement learning are improved [9], [22], [24], [26], [30]. Since reinforcement learning acquires exhaustive rules for selecting appropriate actions to an intended problem and then it holds a sufficient number of rules to deal with possible states of the problem, it works efficiently for relatively small-scale problems. However, for large scale problems or problems with many aliased states, it may perform poorly because of explosive growth in the number of rules and the use of memories.

In classifier systems the idea of reinforcement learning is implemented in a sense that Q-learning-like payoff is computed, and classifier systems are extended so as to cope with non-Markov environments or POMDPs [1], [10], [17], [18], [20], [28]. An agent in a classifier system holds rules in if-then type called classifiers, and it employs an action specified in a classifier such that the condition of the classifier matches the information from the environment. In particular, *don't care* denoted by # is introduced in the condition part of classifiers, and conditions corresponding to # match all states. By this capability the rules represented by classifiers are generalized, that is, the agent acquires the ability to hold classifiers matching multiple different states of the environment. Compared to reinforcement learning, it is thought that the number of rules is smaller and memories are efficiently used in classifier systems, and genetic algorithms can be applied to a set of rules represented in if-then format without difficulty for evolving the rule set suitably. From these features of classifier systems, it is adequate to apply them to problems in non-Markov environments or POMDPs.

In this paper, we develop a learning classifier system for non-Markov environments or POMDPs where a mechanism for controlling the internal register is separated from classifiers and aliased states are identified by detecting fluctuation of the payoffs received by classifiers. We call the proposed system XCSAT (XCS with an internal Action Table) because it is characterized by an internal action table which is a set of rules for identifying aliased states. In XCSAT, after detecting the fluctuation of payoffs which means the existence of aliased states, the environmental information and the corresponding update of the internal register are recorded in the internal action table as a rule for updating the internal register. By controlling the internal register through the information from the internal action table, more efficient and stable performance can be expected in XCSAT.

The remainder of this paper is organized as follows. After describing non-Markov environments in section 2, we mention the properties of XCSM and XCSMH developed by Lanzi and Wilson [20] in section 3. In section 4, we propose a learning classifier system with the internal action table, XCSAT, in which aliased states are identified by detecting fluctuation of payoffs and referring to the internal action table. The experimental result of XCSAT is shown, compared with XCSM and XCSMH in section 5, and finally, section 6 concludes with some comments.

II. Non-Markov Environments

Markov environments have memoryless property, that is, in Markov environments the probability of being in a given state depends on the current state and action but not on any past states or actions, and environments without such property are said to be non-Markov environments. Learning classifier systems for non-Markov environments have been proposed, and to evaluate their performances, woods environments which are grid-like non-Markov environments are used [1], [17], [18], [20], [28].

First of all, to understand that it is difficult for learning classifier systems which are not developed specially for non-Markov environments to find an optimal policy in non-Markov environments, we illustrate actions of an agent in a simple woods environment termed **Woods100** [18], which is shown in Fig. 1.

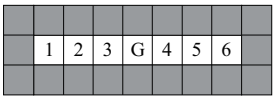


Fig. 1: Woods environment: **Woods100**

In **Woods100**, there are 7 cells which are cell 1 to cell 6 and cell G meaning the goal, and the 7 cells are surrounded by the walls. Although the agent can generally move to 8 possible directions (N, S, E, W, NE, SE, NW, and SW) in a woods environment. In **Woods100**, the agent in any of the 7 cells can move only to W (left) or E (right). The agent starts from cell 1 or cell 6, and it tries to reach cell G. Since the agent moves either to left or to right, the condition part of classifiers deals with states of cells located on only both sides of the agent.

In Table I, classifiers which lead the agent in each cell to the goal are enumerated, and "w" indicates the wall, "c"

TABLE I: Classifiers corresponding to **Woods100** and cells

condition		action	cell
left	right	direction	
w	c	right	1
c	w	left	6
c	c	right	2, 5
c	c	left	2, 5
c	G	right	3
G	c	left	4

(w: wall, c: corridor, G: goal)

indicates the corridor, and "G" indicates the goal in the first and second columns. Take a classifier in the first row of Table I as an example. The first classifier means "If [{*left*: w} and {*right*: c}] then [*external action*: move right]." Therefore, one finds that this classifier should be selected in cell 1, which is given in the rightmost column of Table I.

Since there are 6 available cells except for the goal cell, an optimal policy can be described by 6 classifiers as shown in Table I, and there are two classifiers with the condition parts matching both of the environmental states corresponding to cells 2 and 5. Although the environmental information perceived by the agent in cell 2 is the same as that in cell 5, optimal actions in the two cells are different. From this fact, these two cells are aliased states for the agent, and the agent informed of only the environmental information at the current period cannot find the optimal policy. Thus, it follows that a woods environment such as **Woods100** is one of non-Markov environments.

III. Learning Classifier Systems with Internal Memory

To cope with environmental aliasing, Lanzi and Wilson [20] develop XCSM (XCS with internal memory) which is an extension of XCS. In XCSM, a condition for the internal register and an action for controlling the internal register are added to the condition part and the action part of a classifier, respectively.

TABLE II: Classifiers of XCSM and the related information

no.	condition part			action part		cell	payoff
	left	right	register	direction	revised register		
a	w	c	0	right	0	1	$\gamma^2 R$
b	c	c	0	right	0	2, 5	γR
c	c	G	0	right	0	3	R
d	c	w	0	left	1	6	$\gamma^2 R$
e	c	c	1	left	1	2, 5	γR
f	G	c	1	left	1	4	R
g	w	c	1	right	0	1	$\gamma^2 R$
h	c	G	1	right	0	3	R
i	c	w	1	left	1	6	$\gamma^2 R$
j	G	c	0	left	1	4	R

In Table II, we give an example of an optimal policy in XCSM to **Woods100**, which can be obtained after enough learning process. Since it is necessary for XCSM to discriminate the two aliased states in **Woods100**, only the size of two is required for the internal register. Let the initial value of the internal register be 0. To utilize the internal register, the value of the internal register and its new value to be updated are added in the condition part and the action part of classifiers, respectively, as seen in Table II. For example, in classifier *a* given in Table II, a condition on the internal register "if [*internal register*: 0] (if the internal register is 0)" is given in addition to a condition on the environmental information "if [{*left*: w} and {*right*: c}] (if the left-side cell is the wall and the right-side cell is the corridor)," and an action to the internal register "[*internal action*: set 0] (set 0 in the internal register)" is also given in addition to an action to the environment "[*external action*: move right]." Although the information from the environment when being in cell 2 is the same as that in cell 5 in XCS, since in XSCM the value of the internal register is changed from 0 to 1 by classifier *d* used at cell 6 which is located on the right side of cell 5 but it is not changed to cell 2, cells 2 and 5 can be distinguished. It should be noted that a set of classifiers shown in Table II is an optimal policy, but there exist other sets of optimal policies such as an optimal policy of the reversed procedure.

A classifier in XCSM has the same parameter set as those of XCS: the prediction p, the prediction error ϵ, and the fitness F. The prediction p is a payoff that the system expects if the condition of the classifier conforms with the environmental state and the action of the classifier is performed. The prediction error ϵ estimates an error of the prediction p by using the Q-learning-like payoff. The fitness F means the accuracy of the prediction p and it is a function of the prediction error ϵ. Moreover, the learning process of XCSM is similar to that of XCS, and it is slightly modified for introducing the internal register.

Although XCSM can find an optimal policy as seen in Table II, depending on environments, the sequence of actions may not converge because actions to the environment and to the internal register are determined according to received rewards. We illustrate this difficulty by using **Woods100**. Let R denote the reward from the environment when the agent reaches the goal, and any reward is not paid by arriving at the other cells. When the sequence of actions converges, the payoffs received by the classifiers are shown in the rightmost column of Table 2, where γ is a discount factor.

Assume that the following classifier *a'* is included in the system in addition to the set of classifiers given in Table II:

(*a'*) If [{*left*: w} and {*right*: c}] & [*internal register*: 0] then [*external action*: move right] & [*internal action*: set 1].

Classifier *a'* is the same as classifier *a* except for the internal action, which means the value of the internal register to be updated, in the action part of classifiers. Although, as a matter of course, using classifier *a'* instead of classifier *a* is not optimal, the payoff of classifier *a'* is $\gamma^2 R$ which is the same as that of classifier *a* if classifiers *e* and *b'* are used, where *b'* is the same as classifier *b* except for the internal action. Thus, since the fitness F is a function of the payoff, it is possible that classifier *a'* is substituted for classifier *a*, and therefore it is difficult to generate an optimal policy stably. Beside, it should be noted that when the size of the internal register becomes larger, the performance of XCSM grows worse due to increase of the search space.

To improve the performance of XCSM, XCSMH is also

proposed as an extension of XCSM, and the following modifications are remarked.

(i) The value of the internal register is changed only if the environmental information perceived by the agent changes as a result of the executed action to the environment.
(ii) The actions to the environment and to the internal register are performed in a stepwise fashion. After the value of the internal register is determined by the greedy method, the action to the environment is selected by the ϵ-greedy method.

For example, by reason of (i), it is not possible that the direction to move to a cell of wall is chosen and the value of the internal register is updated at the same time. The modification of (ii) facilitates the combination of actions to the environment and treatment of the internal register, and then it is thought that the performance is improved.

However, XCSMH does not resolve the above mentioned difficulty essentially, and we need some solution to effectively manage the internal register. In this paper, focusing on fluctuation of payoffs of classifiers used in aliased states, we propose an effective learning classifier system with an internal action table providing stable performances by separating the control of the internal register from the action part of classifiers.

IV. Classifier system with an internal action table

As we pointed out before, in non-Markov environments or POMDPs, although XCSM can find an optimal policy, depending on environments, its performance is not always stable because actions to the environment and to the internal register are determined according to received rewards. We will show this fact by some computational experiments in the following section. In this paper, we develop a learning classifier system called XCSAT (XCS with an internal Action Table) for non-Markov environments or POMDPs where controlling the internal registers is separated from classifiers and aliased positions or states are identified by detecting the fluctuation of the payoffs received by classifiers. In XCSAT, after detecting the fluctuation of payoffs, the corresponding environmental information and the updated value of the internal register are recorded into the internal action table as a rule for updating the internal register. By introducing the above mentioned two features simultaneously, XCSAT works efficiently for non-Markov environments or POMDPs.

To check whether or not a position that the agent have arrived is an aliased one, XCSAT focuses on the fluctuation of payoffs received by classifiers. The maximum and the minimum payoffs are recorded together with the corresponding periods of time. If the difference between the maximum and the minimum is larger than the threshold after a given amount of periods had elapsed, XCSAT judges that the payoffs of the classifier fluctuate.

If the payoffs of the classifier executed at the present moment, say period t, does not fluctuate and the payoff fluctuation is observed at period $t-1$, XCSAT judges that the environment at period $t-1$ is an aliased state. To utilize the information about such aliased states, the system records the external state, the internal state and the internal action which are observed and selected at period $t-2$ into the internal action table. By referring to the internal action table with the information about the aliased states each period, XCSAT can identify each aliased state and select an appropriate action for the aliased state.

A. Rule representation and the internal action table

In the proposed method, states of the environment are identified by observing the payoffs received by classifies and referring to the internal action table. To do so, the system stores rules for updating the internal register in the internal action table. Unlike XCSM and XSCMH, XCSAT does not use classifiers to control the internal register, but to this end it uses the internal action table in which the history of use of the internal register is stored.

To describe the learning procedure of XCSAT, we define the following technical terms. Let "an *external state*" be an environmental state, "an *internal state*" be the value of the internal register, "an *external action*" be an action taken by the agent to the environment, and "an *internal action*" be the value of the internal register to be updated.

Using these terms, we represent a classifier in XCSM or XCSMH by the following if-then rule:

If [*external state*] & [*internal state*]
then [*external action*] & [*internal action*].

It should be noted that an *internal action* is specified in the action part of a classifier in XCSM or XCSMH. In contrast, a classifier in XCSAT is expressed as

If [*external state*] & [*internal state*] & [*internal action*]
then [*external action*],

where, in the action part, there does not exist an *internal action*, but it is in the condition part. The *internal action* in the condition part is utilized to update the parameters of a classifier when the classifier is selected to activate to the environment. Apart from classifiers, rules for updating the internal register are stored in the internal action table in the following form:

If [*external state*] & [*internal state*] then [*internal action*].

If the environmental state and the value of the internal register coincide with the values of the *external state* and *internal state* of a rule in the internal action table, respectively, the value of the internal register is updated by using the value of the *internal action* of the rule in the internal action table for updating the internal register. Since the value of the internal register is determined as just described, classifiers in XCSAT have no information about *internal actions* in the action part.

B. Update and usage of the internal action table

Using **Woods100** shown in Fig. 1 and Table II, we illustrate the fluctuation of payoffs received by classifiers used in aliased states. Assume that the sequence of actions of the agent converges through enough learning process. A payoff of classifier *c* which is used at cell 3 and leads to the goal, cell *G*, and that of classifier *f* which is used at cell 4 and also leads to

cell *G* are the same value R. If classifier *b* is instantly used at cell 2 and then classifier *c* is used at cell 3, classifier *b* receives the payoff of γR, where γ is a discount factor. However, if, after classifier *e*, which should be used ideally at cell 5, is used at cell 2, the agent returns to cell 2, classifier *b* is used at cell 2 and then finally classifier *c* is used at cell 3, then classifier *b* receives the payoff of $\gamma^3 R$. If classifier *e* is used at cell 2 repeatedly, the payoff of classifier *b* becomes smaller. Thus, the payoff of classifier *b* ranges from $\gamma^3 R$ to some small value, and as for classifier *e*, a similar fluctuation of the payoff can be observed. Moreover, since the payoffs of classifiers *a* and *d*, which should be used ideally at cells 1 and 6, respectively, are calculated from those of classifiers *b* and *e*, they also fluctuate. From this observation, if the payoff of a classifier used at a certain cell, say cell *x*, fluctuates and cell *x* is adjacent to a cell such that the payoff of the corresponding classifier does not fluctuate, it can be inferred that an environmental state when being in cell *x* is an aliased state. To utilize such information, rules for identifying aliased states are stored in the internal action table.

Although, in XCSAT, an external action which is an action taken by the agent to the environment is selected among classifiers matching the environmental state, an internal action for updating the internal register is determined by finding a rule conforming with the external state and the internal state in the internal action table. As we mentioned above, the form of rules in the internal action table is "If [*external state*: $\cdots$] & [*internal state*: $\cdots$] then [*internal action*: $\cdots$]," and a rule conforming with the external state and the internal state perceived by the agent is searched in the internal action table. An internal action of the rule selected from the internal action table is performed. By doing so, XCSAT can identify aliased states and select appropriate external actions. Eventually, the fluctuation of classifiers' payoffs disappears and an optimal policy can be found. If the payoff fluctuation of classifiers is still observed, it follows that there exist aliased states which are not identified by the system yet.

We demonstrate a process of updating the internal register by using **Woods 100** shown in Fig. 1. Examples of classifiers and the internal action table of XCSAT are given in Tables III and IV. In the course of repetition of trials in **Woods100**, suppose that the fluctuation of payoffs received by a classifier is observed and it is revealed that an environmental state when being in cell 2 is an aliased state. At this point, a rule for the internal register "If [{*left*: w} and {*right*: c}] & [*internal register*: 0] then [*internal action*: set 1]" is recorded in the internal action table, and this rule for the internal register corresponds to cell 1. Moreover, at the same time, a new classifier corresponding to the same external state and the internal state that the value of the internal register is 1 ([*internal register*: 1]) is added into the system. More specifically, the following classifier is generated: If [{*left*: w} and {*right*: c}] & [*internal register*: 1] & [*internal action*: set 1] then [*external action*: move right].

In general, if XCSAT finds an aliased position or state, a rule for identifying the aliased state is recorded in the internal action table. By referring to the internal action table, XCSAT can efficiently distinguish positions of the agent. More precisely, if the payoffs received by the classifier executed at period t does not fluctuate and the payoff fluctuation at period

TABLE III: Classifiers of XCSAT for woods100 and the related information

no.	condition part: external state left	condition part: external state right	condition part: internal register	condition part: internal action	action part: external action	cell	payoff
a	w	c	#	#	right	1	$\gamma^2 R$
b	c	c	1	1	right	2, 5	γR
c	c	G	#	#	right	3	R
d	c	w	#	#	left	6	$\gamma^2 R$
e	c	c	0	0	left	2, 5	γR
f	G	c	#	#	left	4	R

TABLE IV: Internal action table of XCSAT for woods100

no.	condition part: external state left	condition part: external state right	condition part: internal register	action part: internal action	cell
a	w	c	0	1	1
b	c	w	1	0	6

$t-1$ is observed, it is judged that the environment at period $t-1$ is an aliased state. To execute this procedure successfully, XCSAT records the external state, the internal state and the internal action which are observed and selected at period $t-2$ in the internal action table. Thereafter, by referring to the internal action table, it acquires ability to distinguish such states of the environment. The data insertion of the internal action table and the generation of the corresponding classifier are summarized as follows.

Step 1 Refer to the internal action table, and then if XCSAT finds a rule with the condition matching the current environment and the internal register, update the internal register to the value specified by the rule.

Step 2 Execute an action specified by a selected classifier.

Step 3 If the payoff fluctuation is observed, set the flag for update on and return to Step 1. Otherwise, go to Step 4.

Step 4 If the flag is on, go to Step 5. Otherwise, return to Step 1.

Step 5 Add the information of the external state, the internal state and the internal action which are observed and selected at the period before last in the form of

If [*external state*: $\cdots$] & [*internal state*: $\cdots$] then [*internal action*: $\cdots$],

into the internal action table. Moreover, add a new classifier consisting of the information from the environment, the value of the internal register, the updated value of the internal register and the executed external action at the last period in the form of

If [*external state*: $\cdots$] & [*internal state*: $\cdots$] & [*internal action*: $\cdots$] then [*external action*: $\cdots$].

Then, after setting the flag off, return to Step 1.

Some explanatory remarks on this procedure follows.

- After the elapse od the given periods, say 1000 periods, XCSAT starts to refer the internal action table because it needs enough learning time for external environments except aliased states.

- In Step 3, if the difference between the maximum and the minimum of the payoffs received by the selected classifier is larger than the threshold, XCSAT concludes that the payoff fluctuation of the classifier is observed.
- The rules for the internal register are not deleted unless the number of rules exceeds the capacity for the internal action table, and if it exceeds the capacity, the rule with the lowest use is replaced with a new rule for the internal register.
- If two or more aliased states adjoin and there exist multiple such adjoining aliased states, the fluctuation of the payoffs could not be always suppressed. When the payoff fluctuation cannot be suppressed within a given amount of periods after the last update of the internal action table, even if the condition of Step 4 is not satisfied, with a given probability a new rule for the internal register is added into the internal action table.

C. Algorithm of XCSAT

The algorithm of XCSAT is summarized as follows.

Step 1 After perceiving the current external and internal states, XCSAT finds all the classifiers satisfying these conditions. A set of these classifiers are called the match set $[M]$.

Step 2* If a rule matching the perceived external and internal states is found in the internal action table, an internal action specified by the rule is executed. Otherwise, reset the internal register, i.e., set 0 at the internal register.

Step 3 For classifiers with the executed internal action in $[M]$, a prediction array is calculated by using the prediction and the fitness. From the prediction array, an external action is determined by the greedy or the ϵ-greedy method. A set of classifiers with the selected external action in $[M]$ is called the action set $[A]$.

Step 4 After updating the parameters of each classifier in $[A]$, the genetic operations of reproduction, crossover and mutation are performed to the condition part of the classifiers.

Step 5* If the condition based on the payoff fluctuation for updating the internal action table is satisfied, the corresponding external state, internal state, and internal action are recorded in the internal action table.

Step 6 If the agent reaches the terminal position, the algorithm stops. Otherwise, go to Step 1,

It should be noted that as mentioned in the previous subsection, to find appropriate actions for non-aliased states, for the given initial ceratin periods, XCSAT does not refer the internal action table, and therefore Steps 2 and 5 marked with an asterisk, which involve reference and update to the internal action table, are skipped in the initial certain periods, namely it performs the same procedure as that of XCS.

Let the ith classifier be denoted by cl_i. Similarly to those of XCS [3], [20], the main parameters of classifier cl_i are the prediction $cl_i.p$, the prediction error $cl_i.\epsilon$, and the fitness $cl_i.F$. These parameters are updated based on the payoff P received by a classifier and the other parameters. A classifier in the action set $[A]$ receives the following Q-learning-like payoff:

$$P = \begin{cases} R, & \text{when reaching the termination position} \\ P_{-1} + \gamma \max PA, & \text{otherwise,} \end{cases} \tag{1}$$

where R is the reward from the environment, P_{-1} is the payoff at the previous period, PA is the prediction array at the current period, and γ is a discount factor. For a given external action a_i, an element of PA is calculated as follows:

$$PA(a_i) = \sum_{cl_k \in [M]_{\hat{m},a_i}} cl_k.p \cdot cl_k.F \Big/ \sum_{cl_k \in [M]_{\hat{m},a_i}} cl_k.F, \tag{2}$$

where $[M]_{\hat{m},a_i}$ is a set of classifiers in $[M]$ such that an *internal action* is the executed internal action $\hat{m}$ and an *external action* is a_i. In Step 3, by using the prediction array PA, an external action is determined.

The prediction $cl_i.p$ and the prediction error $cl_i.\epsilon$ are updated as follows:

$$cl_i.p = cl_i.p + \beta(P - cl_i.p), \tag{3}$$

$$cl_i.\epsilon = cl_i.\epsilon + \beta(|P - cl_i.p| - cl_i.\epsilon), \tag{4}$$

where β is the learning rate. The smaller the prediction error $cl_i.\epsilon$, the larger the fitness $cl_i.F$ becomes. To this end, the accuracy $cl_i.\kappa$ is defined as

$$cl_i.\kappa = \begin{cases} 1 & \text{if } cl_i.\epsilon < \epsilon_0 \\ \alpha \left(\frac{cl_i.\epsilon}{\epsilon_0} \right)^{-\nu} & \text{otherwise,} \end{cases} \tag{5}$$

where α, ν, and ϵ_0 are parameters, the fitness $cl_i.F$ is calculated as follows:

$$cl_i.F = cl_i.F + \beta(cl_i.\kappa' - cl_i.F), \tag{6}$$

where $cl_i.\kappa' = cl_i.\kappa / \sum_{cl_k \in [A]} cl_k.\kappa$.

As for the genetic operations described in Step 4, if the average elapsed time periods of classifiers in the action set $[A]$ after the last genetic operations for them is larger than a given time period θ_{GA} in advance, the genetic algorithm are executed to the parts of classifiers describing the external conditions. Overgeneral rules in XCSAT are removed in the same way as in XCS. That is, since the prediction errors of overgeneralized classifiers become large and then their fitness in the genetic algorithm described in Step 4 degrades, such classifiers are not reproduced eventually. Two classifiers are chosen by using the roulette wheel selection, and the one-point crossover is applied to them. If a gene chosen for mutation is #, which means "don't care," the perceived external state is filled in the gene. Otherwise, it is exchanged for #.

To judge the fluctuation of the payoffs in Step 5, the maximal payoff $cl_i.p_{\max}$ and the minimal payoff $cl_i.p_{\min}$ are recorded together with the corresponding periods $cl_i.t_{\max}$ and $cl_i.t_{\min}$. Let P be the payoff of classifier cl_i. If $P > cl_i.p_{\max}$, the maximal payoff $cl_i.p_{\max}$ is updated, and similarly if $P < cl_i.p_{\min}$, the minimal payoff $cl_i.p_{\min}$ is updated. Let $cl_i.exp$ be the number of updating, and θ_p and θ_t be parameters. If $cl_i.p_{\max} - cl_i.p_{\min} < \theta_p$ and $cl_i.exp > \theta_t$, XCSAT judges that the payoff of classifier cl_i does not fluctuate. Otherwise, it

judges that the payoff of the classifier fluctuates. Furthermore, let θ_r be a parameter. If the elapsed time periods after the last update of either $cl_i.p_{\max}$ or $cl_i.p_{\min}$ is larger than θ_r, the not yet updated parameter and the update counter $cl_i.exp$ are initialized.

V. Computational experiment

To demonstrate the effectiveness of XCSAT, we perform a computational experiment by using a woods environment **Woods101$\frac{1}{2}$**, and compare XCSAT with XCSM and XCSMH. Furthermore, with another eight woods environments [1], [17], [20], [21], [28], we examine the performance of XCSAT.

A. *Woods101$\frac{1}{2}$*

In XCSAT, the agent perceives substances of the adjacent eight cells (N, S, E, W, NE, SE, NW, and SW), and it distinguishes among "wall," "corridor," and "the goal" of substance of each of the cells. As seen in Fig. 2, **Woods101$\frac{1}{2}$** is a separated symmetric woods environment, and the agent tries to move from any cell labeled as S to one of the cells labeled as G in the shortest possible route.

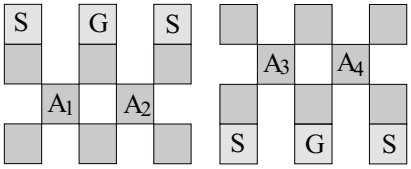


Fig. 2: Woods environment: **Woods101$\frac{1}{2}$**

Since, in the cells labeled as A_1, A_2, A_3, and A_4, substances of the eight cells that the agent perceives are the same, the agent cannot distinguish these states of the environment. In cells A_1, A_2, A_3, and A_4, "move upper right," "move upper left," "move lower right," and "move lower left" are appropriate actions, respectively.

An episode is defined as a process from starting at cell S to reaching cell G. Let one trial be 8000 episodes; the periods until episode 6000 are served for exploring or learning, and the remaining 2000 episodes are used for test of the performance. While the ϵ-greedy method which includes stochastic selection of actions is employed in the learning periods, the greedy method in which an action with the largest prediction is chosen with certainty is used in the test periods. To examine the performances of XSCAT, XCSM and XCSMH, data from the last 1000 episodes are used for each trial, and their performances are evaluated by the average of 30 trials. The parameters used in the computational experiment are shown in Table V. In the experiment, we use XCSM and XCSMH programs of our own making according to the procedure given in Lanzi and Wilson [20]. The sizes of the internal registers in three programs, XCSAT, XCSM and XCSMH, are all 4 for the seven problems in sections 5.1 and 5.2, and they are 6 and 8 for the two problems, **Lab1** and **LargeMaze**, in section 5.3., respectively.

In Fig. 3 and Fig. 4, we compare the performances of XCSAT, XCSM, and XCSMH, varying the exploration rate ϵ in the ϵ-greedy method in the learning periods. The rate of convergence in Fig. 3 is the percentage of success in the 30 trials, and the success means that the agent exactly takes a shortest route and reaches the goal in the last 1000 episodes in the test periods. Aside from this, the rate of convergence in Fig. 4 is the percentage in the 30 trials that the agent takes the same fixed route including the shortest route in the last 1000 episodes. Therefore, the term "the convergence" means that the agent takes the same route for a given starting point in the last 1000 episodes.

TABLE V: Parameters

learning rate	$\beta = 0.2$	discount factor	$\gamma = 0.71$
periods for GA	$\theta_{GA} = 25$	crossover probability	$p_c = 0.75$
mutation probability	$p_m = 0.025$	accuracy parameters	$\alpha = 0.1$, $\nu = 5$
payoff range	$\theta_p = 5$	updating counter	$\theta_t = 30$
periods for updating	$\theta_r = 10$		

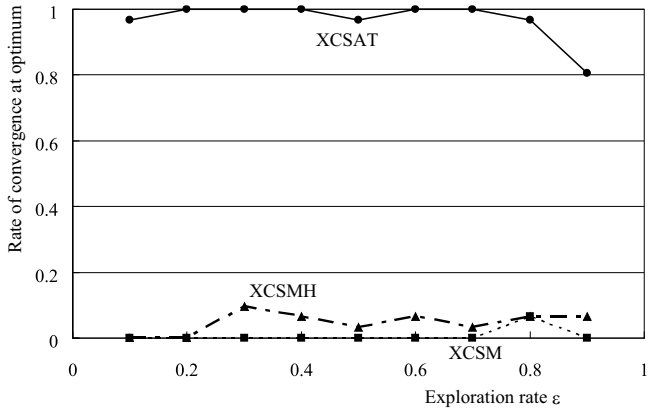


Fig. 3: Convergence on the shortest routes

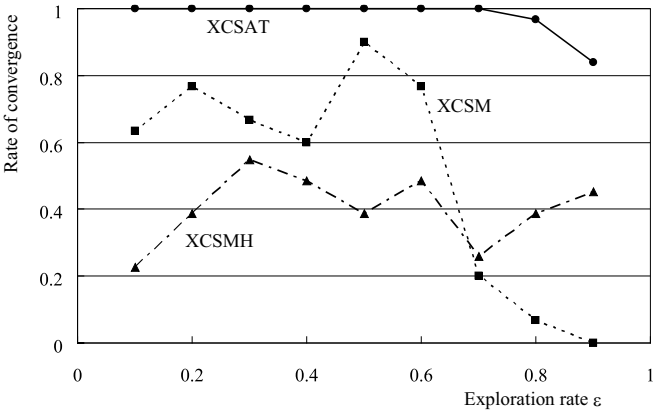


Fig. 4: Convergence on fixed routes

As seen in Fig. 3 and Fig. 4, while XCSM and XCSMH can hardly find the shortest routes, the agent in XCSAT pursues the shortest routes with great accuracy. Although the rate of convergence on the shortest routes in XCSMH is slightly larger than that of XCSM, for the convergence on another fixed route, the rate of XCSM is larger than that of XCSMH. The number

of steps taken from the starts to the goals in XCSM and XCSMH is larger than 12 on an average, and thus it follows that the routes taken by the agent in these systems converge some fixed routes including unnecessary actions because the number of steps for the shortest routes is 4.

B. Performance verification

We continue to examine the performance of XCSAT by using other 6 woods environments which are depicted in Fig. 6 in the appendix, and the summary data of them are given in Table VI [17], [20], [21], [28].

TABLE VI: Woods environments for the computational experiment

woods environment	number of all states	number of aliased states
woods101	11	4
woods102	28	10
maze7	10	2
mazeF4	11	2
maze10	19	13
Littman57	15	8

In these woods environments, starting cells are randomly chosen from among non-goal cells. We evaluate the performances by measuring the number of steps taken from the starts to the goals. In the experiment, if the number of actions taken by the agent is larger than 10000 and the agent still does not reach the goal, the current episode terminates and the next episode begins with a new starting cell. The exploration rate is fixed at $\epsilon = 0.5$. The other experimental conditions and the parameters are the same as those in the computational experiment for **Woods101$\frac{1}{2}$** given in section V-A.

The result of the computational experiment is given in Fig. 5 and Table VII. The performances of the three systems XCSAT, XCSM and XSCMH are compared on the basis of the data of the last 1000 episodes for the 30 trials. The term *best* in Fig. 5 and Table VII means the minimum among the results of the 30 trials where the result of each trial is the average of the last 1000 episodes. Therefore, we note that the *best* is not always the optimum. The terms *mean* and *worst* also mean the average of the 30 trials and the maximum among the 30 trials, respectively.

In Fig. 5, *best*, *mean*, *worst*, and the range of the steps taken by the agent from the starts to the goals are given graphically. In particular, *mean* is denoted by a circle, *best* and *worst* are denoted by bars, and the range of the steps is represented by vertical lines. In Table VII, the minimal steps among the three systems are emphasized by boldface, and for reference the average steps of the shortest routes are given in the rightmost column. For example, the average of shortest steps of *mazeF4* is calculated by summing up the numbers of the shortest steps to the goal for all cells and dividing the number of cells, i.e.,

$$(4+3+2+1+0+4+5+5+6+7+6)/11 = 3.90.$$

As seen in Table VII, the number of steps of XCSAT for each of the six woods environments is close to the average step of the shortest routes. The *best* of XCSMH is smaller than that of XCSM, and XCSMH provides comparable result to that of XCSAT. However, the *mean* and *worst* of XCSMH

TABLE VII: Results of the computational experiment (steps)

woods environment		XCSAT	XCSMH	XCSM	average of shortest steps
woods101	mean	**2.70**	22.38	3.19	2.45
	best	**2.62**	2.64	2.64	
	worst	**2.85**	295.92	4.11	
woods102	mean	**3.28**	6.29	6.73	2.57
	best	**3.00**	4.23	4.93	
	worst	**3.73**	14.63	12.45	
maze7	mean	**4.19**	44.51	38.57	3.70
	best	**3.90**	**3.90**	4.93	
	worst	**7.75**	392.63	126.26	
mazeF4	mean	**4.40**	116.89	35.72	3.90
	best	**4.09**	4.11	4.18	
	worst	**5.54**	1171.19	133.47	
maze10	mean	**6.51**	12.08	46.45	4.32
	best	**5.70**	6.39	8.20	
	worst	**8.54**	61.76	173.39	
Littman57	mean	4.19	**3.89**	6.85	3.47
	best	**3.47**	**3.47**	5.52	
	worst	5.99	**5.14**	9.87	

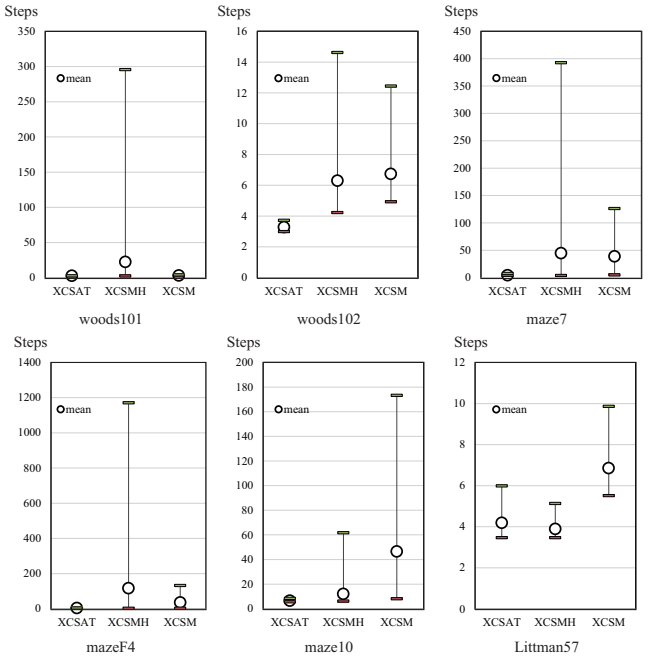


Fig. 5: Results of the computational experiment (steps)

are obviously larger than those of XCSAT, and in some woods environments the *mean* and *worst* of XCSMH sometimes are larger than those of XCSM. This means that the learning performance of XCSMH is not stable, and we consider that this difficulty is attributed to the problem described in section III. In contrast, XCSAT works well in finding the shorter routes to the goal, and then the performance of XCSAT is stable as seen Table VII. As we discussed in section V-A, also from the viewpoint of the convergence given in Fig. 3 and 4, the performance of XCSAT is more stable than those of XCSM and XCSMH. In general, the performance of XCSAT is superior to XCSM and XCSMH except for **Littman57**. In the experiment for **Littman57**, XCSMH shows the best performance but the performance of XCSAT is also good. Both of XCSAT and XCSMH find the shortest routes, and the

difference between the mean steps of them is only 0.3 steps.

As for Littman57, which is composed of 15 states including 8 aliased states, XCSMH works slightly better than XCSAT, and the performance of XCSM is also reasonable. For maze7 or mazeF4, however, XCSMH and XCSM operate inefficiently despite the fact that maze7 or mazeF4 is composed of 10 or 11 states and there are only two aliased states in both of them. This performance is thought to be due to the property of the optimal actions. That is, optimal actions are the same in the aliased states of Littman57, while they are different actions in those of maze7 or mazeF4. XCSAT works well in both problems because it finds an appropriate action efficiently by referring the internal action table.

C. Performance and adaptability for larger problems

The average numbers of all states and aliased states of the woods environments dealt with in section V-B are 15.7 and 6.5, respectively. To examine the effectiveness of XCSAT to lager problems, we use a woods environment **Lab1** [1] which is 5 times as large as the woods environments in section V-B, and we also provide a new woods environment **LargeMaze** which is 10 times as large as them. These woods environments are depicted in Fig. 7 of the appendix.

To cope with the larger problems, we revise the condition for judging the payoff fluctuation of a classifier. The condition given in section IV is that $cl_i.p_{\max} - cl_i.p_{\min} < \theta_p$ and $cl_i.exp > \theta_t$, and if this condition is satisfied, XCSAT judges that the payoffs of the classifier does not fluctuate. Since the difference between $cl_i.p_{\max}$ and $cl_i.p_{\min}$ depends on the size of a problem and the payoff of a classifier decreases according to the discount factor γ every period of time, by using the discount factor γ, we employ a modified condition $\gamma cl_i.p_{\max} < cl_i.p_{\min}$ instead of $cl_i.p_{\max} - cl_i.p_{\min} < \theta_p$. The remaining procedure for judging the payoff fluctuation is the same as before, and the value of γ is set at $\gamma = 0.9$ due to increase of the problem size. The system with the revised condition for judging the payoff fluctuation is denoted by XCSATγ. The number of trials is 20. The other experimental conditions and the parameters are the same as those in the computational experiment for the six woods environments given in section V-B.

In this computational experiment, the performances are evaluated by the average of 20 trials, and we define trials to be valid for measurement as follows: (i) the average steps from the start to the goal is not larger than 100; (ii) a trial, which consists of 8000 episodes, finishes in 5 hours or less.

No trial of XCSM and XCSMH satisfies the two conditions. In XCSAT and XCSATγ, 13 and 14 trials out of the 20 trials meet the conditions for **Lab1**, respectively, and 5 and 6 trials meet them for **LargeMaze**, respectively. That is, XCSM and XCSMH are no longer workable for larger scale maze problems such as **Lab1** and **LargeMaze**, while XCSAT or XCSATγ properly works in more than a half of the whole trials for **Lab1** and in a quarter of them for **LargeMaze**. From this result of the computational experiment together with the data shown in the previous subsections, it is understood that the proposed learning classifier systems with an internal action table, XCSAT and XCSATγ, demonstrate superior performance, compared to XCSM and XCSMH.

Since in **LargeMaze**, the numbers of all states and aliased states are large and there exist many possible routes from the starts to the goals, compared to **Lab1**, the number of valid trials of **LargeMaze** is smaller that of **Lab1**, and the steps taken by the agent from the starts to the goals in **LargeMaze** are larger than those of **Lab1**. The performances of XCSAT and XCSATγ are summarized in Table VIII in a way similar to Table VII. As seen in Table VIII, the data supports the superiority of the performance of XCSATγ compared to that of XSCAT, and then the modified condition for judging the payoff fluctuation to larger problems is shown to be effective.

VI. Conclusion

In this paper, we develop a learning classifier system with an internal action table (XCSAT) to deal with sequential decision problems in non-Markov environments. In XCSAT, controlling the internal register is separated from classifiers, and aliased states are perceived by detecting fluctuation of the payoffs received by classifiers. After recognizing the existence of aliased states, the environmental information and the corresponding update of the internal register are recorded in the internal action table as a rule for updating the internal register. XCSAT identifies the perceived aliased state by referring to the internal action table.

By performing computational experiments where 9 woods environments are used, we demonstrate the effectiveness of XCSAT. In particular, XCSAT works well for woods environments such that the number of states are about 20 and the fraction of aliased positions is about 30% as used in Lanzi [18] and Lanzi and Wilson [20].

The success probability of learning for the larger problems by the proposed classifier systems (XCSAT and XCSATγ) are not high, therefore further improvement of the classifier system should be a future work.

References

[1] A. Burisov and A. Vasilyev, "Learning classifier systems in autonomous agent control task," Proceedings of the 5th International Conference on Application on Fuzzy Systems and Soft Computing, 36–42, 2002.

[2] L. Bull, J. Sha'Aban, A. Tomlinson, J. D. Addison and B. G. Heydecker, "Towards distributed adaptive control for road traffic junction signals using learning classifier systems," L. Bull (ed.), *Applications of Learning Classifier Systems*, Springer, New York, 279–299, 2004.

[3] M. V. Butz and S. W. Wilson, "An algorithm description of XCS," *Soft Computing*, 6, 144–153, 2002.

[4] Y. J. Cao, N. Ireson, L. Bull and R. Miles, "Design of a traffic junction controller using a classifier system and fuzzy logic," B. Reusch (ed.), *Computational Intelligence Theory and Applications*, Springer, New York, 342–353, 1999.

[5] D. Cliff and S. Ross, "Adding temporary memory to ZCS," *Adaptive Behavior*, 3, 101–150, 1994.

[6] M. Danek and R. E. Smith, "XCS applied to mapping FPGA architectures," *GECCO '02 Proceedings of the Genetic and Evolutionary Computation Conference*, Morgan Kaufmann Publishers Inc., San Francisco, 912–919, 2002.

[7] K. De Jong, "Learning with genetic algorithms: an overview," *Machine Learning*, 3, 121–138, 1988.

[8] M. Dorigo and M. Colombetti, *Robot Shaping: An Experiment in Behavior Engineering*, MIT Press/Bradford Books, Massachusetts, 1998.

[9] F. Doshi-Velez, J. Pineau and N. Roya, "Reinforcement learning with limited reinforcement: using Bayes risk for active learning in POMDPs," *Artificial Intelligence*, 187/188, 115–132, 2012.

TABLE VIII: Performance for larger problems (steps)

woods environment		XCSAT	XCSATγ	average of shortest steps
Lab1	valid trials	13/20	**14/20**	
number of all states: 79	mean	26.08	**20.99**	
number of aliased states: 35	best	15.59	**15.00**	14.50
	worst	61.22	**52.24**	
LargeMaze	valid trials	5/20	**6/20**	
number of all states: 148	mean	44.71	**37.13**	
number of aliased states: 100	best	31.37	**21.53**	14.00
	worst	55.41	**53.03**	

[10] A. Hamzeh, S. Hashemi, A. Sami and A. Rahmani, "A recursive classifier system for partially observable environments," *Fundamenta Informaticae*, 97, 15–40, 2009.

[11] J. H. Holland, *Adaptation in natural and artificial systems*, University of Michigan Press, Michigan, 1975. (reprinted by the MIT Press in 1992)

[12] J. H. Holland, "Adaptation," R. Rosen, F. M. Snell, (eds.), *Progress in Theoretical Biology*, 4, Academic Press, New York, 263–293, 1976.

[13] J. H. Holland, "Properties of the Bucket Brigade," J. J. Grefenstette (ed.), *Proceedings of the 1st International Conference on Genetic Algorithms*, L. Erlbaum Associates, New Jersey, 1–7, 1985.

[14] J. H. Holland and J. S. Reitman, "Cognitive systems based on adaptive algorithms," D. A. Waterman and F. Hayes-Roth (eds.), *Pattern-Directed Inference Systems*, Academic Press, 313-329, 1978.

[15] J. H. Holmes and J. A. Sager, "Rule discovery in epidemiologic surveillance data using EpiXCS: an evolutionary computation approach," S. Miksch, J. Hunter and E. Keravnou (eds.), *Artificial Intelligence in Medicine*, Springer, Berlin, 444–452, 2005.

[16] L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, 4, 237–285, 1996.

[17] P. L. Lanzi, "An analysis of the memory mechanism of XCSM," *Proceedings of the Third Genetic Programming Conference*, Morgan Kaufmann Publishers, San Fransisco, 643–651, 1998.

[18] P. L. Lanzi, "Adaptive agents with reinforcement learning and internal memory," J. A. Meyer, A. Berthoz, D. Floreano, H. Roitblat, and S. W. Wilson (eds.), *From Animals to Animats 6, Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, MIT Press, Cambridge, 333–342, 2000.

[19] P. L. Lanzi, "Learning classifier systems from a reinforcement learning perspective," *Soft Computing*, 6, 162–170, 2002.

[20] P. L. Lanzi and S. W. Wilson, "Toward optimal classifier system performance in non-Markov environments," *Evolutionary Computation*, 8, 393–418, 2000.

[21] M. L. Littman, A. R. Cassandra and L. P. Kaelbling, "Learning policies for partially observable environments: scaling up," M. N. Huhns and M. P. Singh (eds.), *Readings in Agents*, Morgan Kaufmann Publishers, San Francisco, 495–503, 1998.

[22] J. Pineau, G. Gordon and S. Thrun, "Point-based value iteration: An anytime algorithm for POMDPs," mimeo, Carnegie Mellon University, 2003.

[23] R. J. Preen and L. Bull, "Discrete and fuzzy dynamical genetic programming in the XCSF learning classifier system," *Soft Computing* 18, 153–167, 2014.

[24] N. Roy, G. Gordon and S. Thrun, "Finding approximate POMDP solutions through belief compression," *Journal of Artificial Intelligence Research*, 23, 1–40, 2005.

[25] S. Saxon and A. Barry, "XCS and the Monk's problems," P. L. Lanzi, W. Stolzmann and S. W. Wilson (eds.), *Learning Classifier Systems: From Foundations to Applications*, Springer-Verlag, Berlin, 223–242, 2000.

[26] G. Shani, R. I. Brafman and S. E. Shimony, "Model-based online learning of POMDPs," J. Gama, R. Camacho, P. Brazdil, A. Jorge and L. Torgo (eds.), Machine Learning: ECML 2005, Springer-Verlag, Berlin, 353–364, 2005.

[27] W. Stolzmann, "Latent learning in Khepera robots with anticipatory classifier systems," A. Wu (ed.), *Proceedings of the 1999 Genetic and Evolutionary Computation Conference Workshop Program*, Morgan Kaufmann, San Francisco, 290–297, 1999.

[28] W. Stolzmann, "An introduction to anticipatory classifier systems," P. L. Lanzi, W. Stolzmann and S. W. Wilson (eds.), *Learning Classifier Systems: From Foundations to Applications*, Springer-Verlag, Berlin, 175–194, 2000.

[29] W. Stolzmann and M. V. Butz, "Latent learning and action planning in robots with anticipatory classifier systems," P. L. Lanzi, W. Stolzmann and S. W. Wilson (eds.), *Learning Classifier Systems: From Foundations to Applications*, Springer-Verlag, Berlin, 301–320, 2000.

[30] S. D. Whitehead and L. J. Lin, "Reinforcement learning of non-Markov decision processes," *Artificial Intelligence*, 73, 271–306, 1995.

[31] S. W. Wilson, "ZCS: a zeroth level classifier system," *Evolutionary Computation*, 2, 1–18, 1994.

[32] S. W. Wilson, "Classifier fitness based on accuracy," *Evolutionary Computation*, 3, 149–175, 1995.

[33] S. W. Wilson, "Mining oblique data with XCS," P. L. Lanzi, W. Stolzmann and S. W. Wilson (eds.), *Advances in Learning Classifier Systems*, Springer, Berlin, 158–176, 2001.

[34] S. W. Wilson, "Function approximation with a classifier system," *Proceedings of the genetic and evolutionary computation conference (GECCO '01)*, Morgan Kaufmann, San Francisco, 974–981, 2001.

[35] Z. Zang, D. Li, J. Wang and D. Xia, "Learning classifier system with average reward reinforcement learning," *Knowledge-Based Systems*, 40, 58–71, 2013.

APPENDIX

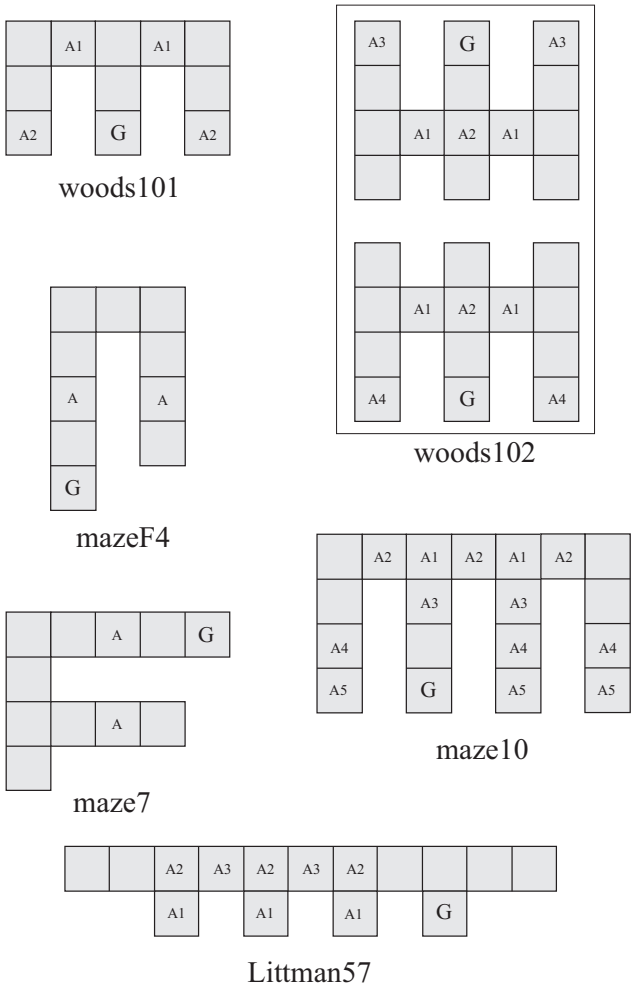


Fig. 6: Woods environments

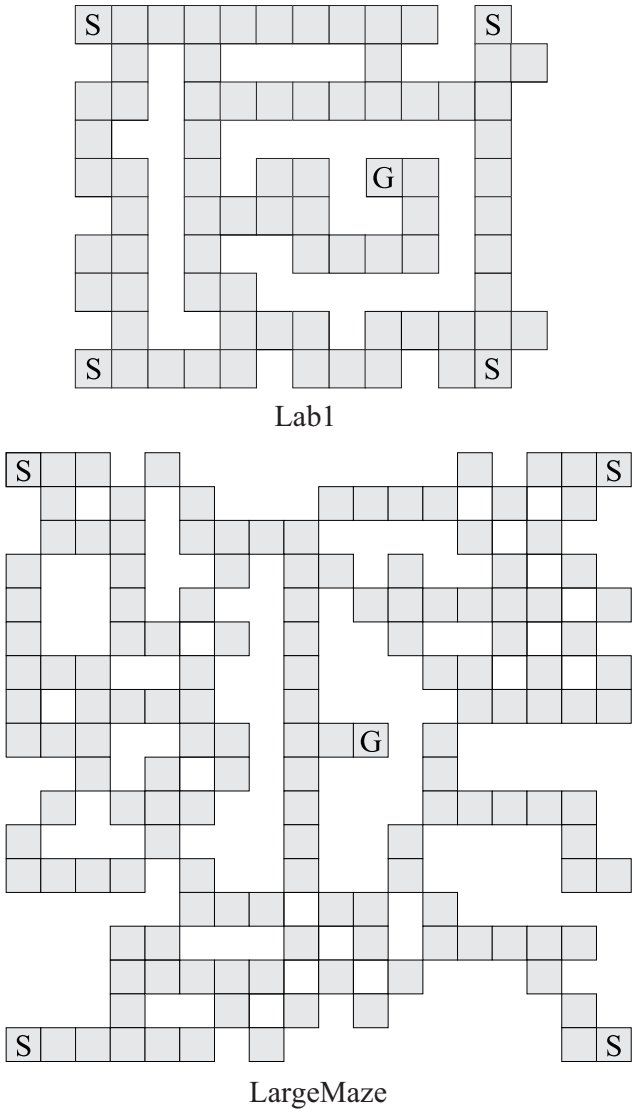


Fig. 7: Large woods environments