IJCSIS Vol. 13 No. 3, March 2015
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA


Cornell University
Library


Cogprints


Google scholar


.docstoc
find and share professional documents


ScientificCommons

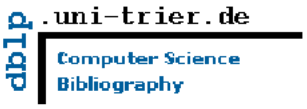

View my documents on
Scribd


BASE
Bielefeld Academic Search Engine


SCIRUS
search engine for science


SciRate.com


CiteSeerx beta


dblp .uni-trier.de
Computer Science
Bibliography


Q·Sensei


DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS


EBSCO
HOST


ProQuest

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2015 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Managing Editor

Over the past several decades, we have witnessed significant research and innovation in several domains including network security, cloud computing and virtualization. The purpose of this edition is to gather novel experimental and theoretical evidence from both industry and academia in the broad areas of Computer Science, ICT & Security and further bring together people who work in the relevant areas. The ***International Journal of Computer Science and Information Security*** *(IJCSIS) promotes research publications which offer significant contribution to the computer science knowledge and which are of high interest to a wide academic/research/practitioner audience. Coverage extends to several main-stream and state of the art branches of computer science, security and related information technology. As a scholarly open access peer-reviewed journal, IJCSIS mission is to provide an outlet for quality research & academic publications. It aims to promote universal access with equal opportunities for international scientific community; to scientific knowledge, and the creation, and dissemination of scientific and technical information.*

IJCSIS archives all publications in major academic/scientific databases. Indexed by the following International agencies and institutions: Google Scholar, CiteSeerX, Cornell's University Library EI, Scopus, DBLP, DOAJ, ProQuest and EBSCO. Moreover, Google Scholar reported increased in number cited papers published in IJCSIS ***(No. of Cited Papers: 554, No. of Citations: 1285, Years: 6th)****. Abstracting/indexing/reviewing process, editorial board and other important information are available online on homepage. By supporting the Open Access policy of distribution of published manuscripts, this journal ensures "free availability on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of [published] articles". We help researchers to succeed by providing high visibility, prestige and efficient publication process.*

IJCSIS editorial board, consisting of international experts, guarantees a rigorous peer-reviewing process. We look forward to your collaboration. For further questions please do not hesitate to contact us at ***ijcsiseditor@gmail.com****.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 13, No. 3, March 2015 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

3. Paper 28021505: Comparative Analysis based Classification of KDD'99 Intrusion Dataset (pp. 14-20)

Shashikant Upadhyay, PG Scholar, Department of Computer Science & Engg, VNSIT, Bhopal, INDIA
Rajni Ranjan Singh, Assistant Professor, Department of Computer Science & Engg, VNSIT, Bhopal, INDIA

Abstract - Intrusion Detection System (IDS) which is used for classifies network traffic information. Mostly traffic information is classified on the basis of normal or anomaly behaviors. In this work an attack specific classification of KDD cup dataset is proposed. Here Naive Bayes, j48, j48 craft, Bayes net and random forest classifiers are explored for the classification. A comparative analysis is carried out to identify best classifiers for each individual attacks based on Precision, recall, f-measures and ROC Curve Area performance criteria's. It is observed that random forest is the best classifier among all used methods. However despite of the high classification of random forest, BayesNet performed better in context of false positive rate of classification. Similarly ROC Curve Area rate of Bayes net is the best among the all used methods.

Keywords- Classification; KDD Cup '99; Naive Bayes; BayesNet J48; J48graft and Random Forest.

4. Paper 28021511: Policy-Based Smart Adaptive Quality of Service for Network Convergence (pp. 21-27)

Ayoub BAHNASSE, Dept. Physic, Lab STIC, Faculty of Sciences EL Jadida, University Chouaïb Doukkali, El Jadida, Morocco
Najib ELKAMOUN, Dept. Physic, Lab STIC, Faculty of Sciences EL Jadida, University Chouaïb Doukkali, El Jadida, Morocco

Abstract — the management of a convergent network has become one of the primordial and challenging tasks. Taking into account the rapid evolution of networks' size and the diversity of deployed applications, and their requirements in terms of quality of service (QoS), it is quite difficult, on the one hand, to ensure through a manual customization of Quality of Service policies a level of performance required by some flows, on the other hand, to automatically and dynamically adapt QoS policies to instantly optimize the resources, in order to guaranty for applications the required bandwidth and to ensure low loss rates , latency and jitter. This paper presents a novel model of the policy-based Smart adaptive Quality of Service management for clients and Internet service providers. The model proposes separation and distribution of: Data, control and management plans, for a quick and effective treatment, as well as automatically and dynamically reallocates bandwidths, reduce the loss rate and minimize the delay and jitter in a converged network.

Keywords- Quality of service, Adaptive QoS, Smart QoS, Policy-based, network convergence.

5. Paper 28021515: Low Footprint Hybrid Finite Field Multiplier for Embedded Cryptography (pp. 28-32)

Sunil D. Bobade, Research Scholar, S.G.B.Amravati University, Amravati, India
Dr. Vijay R. Mankar, Deputy Secretary, MSBTE, Pune Region, Pune, India

Abstract — Finite field multiplier contributes significantly to the area occupancy of cryptoprocessor. In this paper, we propose a novel hybrid finite field multiplier that invokes the more efficient multiplication algorithm. The proposed multiplier switches between two variants of multiplier depending on the size of multiplicands. The Karatsuba multiplier is efficient algorithm ensuring fewer LUTs and stable number of Flip-flops for the smaller bit multiplications, while the other systolic variant ensures fewer LUTs count for the bigger size multiplicands. The proposed hybrid multiplier does the initial recursion using the systolic variant while final small sized multiplications are accomplished using the Karatsuba algorithm. Area analysis report suggests that using a proposed hybrid multiplier instead of just traditional Karatsuba Multiplier, eventually helps in reducing FPGA footprint.

Keywords-Karatsuba Multiplier; Modular Multiplication; Cryptoprocessor; Area complexity.

6. Paper 28021520: Base Station Radiation's Optimization using Two Phase Shifting Dipoles (pp. 33-42)

Safaa BERRA, Laboratory RITM/ESTC, ENSEM,University of Hassan II, Casablanca, Morocco
Mounir Rifi, Laboratory RITM/ESTC, ENSEMUniversity of Hassan II, Casablanca, Morocco

Abstract — The electromagnetic pollution becomes a serious issue, with the increasing number of telecom operators by country. Certainly the technologies differ, in term of naming (BTS, NodeB, E-NodeB), in term of frequencies, but the problem remains the same: how to optimize the radiation of base station? Is there any intelligent system which allows us to reduce the radiation area and to channel it according to the demand? Far from the complicated and expensive networks of antennas, we are going to present in this paper the use of a simple system of antennas to optimize the base station's radiation. We will start this article by presenting some techniques that improve the energy consumption in base station which represent an important ratio of the global energy in a cellular network. Then, we will study the phase shifting effect on the gain and the energy efficiency. This paper aims to use the found results to propose a new process to reduce the radiation of base station. In the third part, we propose a new architecture based on the use of two phased dipoles in the base station to cover in a smart way the subscribers as long as they are present in the cell. We propose a new system that allows us to reduce the radiation area.

Keywords — Smart antenna, Phased array system, Dipole, Base station, Energy consumption, Energy efficiency, Gain, Radiation.

7. Paper 28021529: An Efficient Model to Automatically Find Index in Databases (pp. 43-46)

Mohammad H. Nadimi-Shahraki, Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran
Rezvan Shahriari, Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran
Mohamad Davar Panah Jazi, Department of Computer Engineering, Foulad Institute of Technology, Fouladshahr, Iran

Abstract — The high volume and increasing complexity of data in large organizations requires optimization methods for fast access to data in a database. Choosing the most appropriate index for the database is essential to optimization; the growing need of more complicated indexing makes reliance on current database administrators insufficient. Employing a skilled database administrator to work with complex database management systems is expensive and not an option in many organizations. A more cost-effective approach is use an automated data management system to find the index. Several such approaches have been developed to date. Although there have been many methods based on data mining for finding a proper index, their run time is still unacceptable. Particularly, in very large databases, faster methods for finding proper indexes are needed. In this paper, an efficient model is proposed to automatically find index in databases using maximal frequent patterns. The proposed model is evaluated by conducted benchmark dataset TPC-H to compare with previous methods. The evaluation results show that using the proposed model can decrease the time of finding indexes in databases.

Keywords- Optimization, automatic indexing, maximal frequent patterns

8. Paper 28021530: An Overview and Comparison of Hierarchical P2P-SIP Networks (pp. 47-58)

Abel Diatta & Ibrahima Niang, Departement de Mathematiques et Informatique Laboratoire d'Informatique de Dakar (LID), Universite Cheikh Anta Diop de Dakar , Dakar, Senegal
Mandicou Ba, Universite de Reims Champagne-Ardenne, Laboratoire CReSTIC EA 3804 - Equipe SysCom, Reims, France

Abstract — P2P-SIP networks are emerging distributed communication technology introducing Scalability and Robustness. These networks propose to use a peer-to-peer network for user registration and user location in Session Initiation Protocol (SIP)-based voiceover- IP (VoIP) networks. Nevertheless, P2P-SIP systems are based on a pure flat DHT design which provides a uniform distribution of (equal) peers and resources that assures scalability and load balancing. Recent research works have proposed to organize P2P overlays into hierarchical architecture called

HP2P-SIP. These later are motivated by supporting sophisticated search requirements, separating categories of use (content, personal communications,), scaling, etc. However, these works did not resolve the problem of overloading the bandwidth, which is still a head-case. In this paper, after highlighting the many challenges of hierarchical architectures, we show how to avoid overloading the physical network by acting on the overlay network. To do this, we have a physical network of size N with a number of super nodes (NSN). Our solution shows how many levels k we must build our overlay network so that in the physical network, the number of messages generated by the lookup nodes does not exceed a value x, initially fixed.

Index Terms—HP2P-SIP, Costing, Iterative, Recursive, Semi-recursive, Bandwidth overload

9. Paper 28021531: Unweighted Class Specific Soft Voting based ensemble of Extreme Learning Machine and its variant (pp. 59-65)

Sanyam Shukla, CSE department, M.A.N.I.T, Bhopal, India
R. N. Yadav, ECE department, M.A.N.I.T, Bhopal, India

Abstract — Extreme Learning Machine is a fast real valued single layer feed forward neural network. Its performance fluctuates due to random initialization of weights between input and hidden layer. Voting based Extreme Learning Machine, VELM is a simple majority voting based ensemble of Extreme learning machine which was recently proposed to reduce this performance variation in Extreme Learning Machine. A recently proposed class specific soft voting based Extreme Learning Machine. CSSV-ELM further refines the performance of VELM using class specific soft voting. CSSV-ELM computes the weights assigned to each class of component classifiers using convex optimization technique. It assigns different weights assuming different classifiers perform differently for different classes. This work proposes Un-weighted Class Specific Soft Voting based ensemble, UCSSV-ELM a variants of CSSV-ELM. The proposed variant uses class level soft voting with equal weights assigned to each class of component classifiers. Here all the classifiers are assumed to be equally important. Soft voting is used with the classifiers that have probabilistic outputs. This work evaluates the performance of proposed ensemble using both ELM and a variant of ELM as base classifier. This variant of ELM differs from ELM as it uses sigmoid activation function at output layer to get probabilistic outcome for each class. The result shows that the Un-weighted class specific soft voting based ensemble performs better than majority voting based ensemble.

Keywords—Ensemble Pruning; Extreme learning Machine; soft voting, probabilistic output.

10. Paper 28021541: An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka (pp. 66-74)

Zainal K., Faculty of Science & Technology, Islamic Science University of Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia
Sulaiman N.F., Faculty of Science & Technology, Islamic Science University of Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia
Jali M.Z., Faculty of Science & Technology, Islamic Science University of Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia

Abstract — This paper reported and summarized findings of spam management for Short Message Service (SMS) which consists of classification and clustering of spam using two different tools, namely RapidMiner and Weka. By using the same dataset, which is downloaded from UCI, Machine Learning Repository, various algorithms used in classification and clustering in this simulation has been analysed comparatively. From the simulation, both tools giving the similar results that the same classifiers are the best for SMS spam classification and clustering which are outperformed than other algorithms.

Keywords- SMS spam; RapidMiner; Weka; Naïve Bayesian (NB); Support Vector Machine (SVM); k-Nearest Neighbour (kNN); K-Mean; Cobweb; Hierarchical clustering; spam classification; spam clustering.

11. Paper 28021543: Security Architecture with NAC using Crescent University as Case study (pp. 75-78)

Ayangbekun Oluwafemi J., Department of Information Systems, University of Cape Town, Cape Town, South Africa
Audu Zainab O., Department of Computer Science, Crescent University Abeokuta, Ogun State, Nigeria.

Abstract — A campus network implemented for a university is a medium for effective communication in such environment. The major advantage of having such network is the shared nature of resources and mobile nature of students, teachers and administrators. This research based on the low level of network capacity in Crescent University taking into consideration applying NAC and some selected protocols over both wired and wireless network. Star topology was employed to design the network, a network with centralized server because of the size of the school and the limitations it might encounter if other forms of technologies are deployed, as illustrated in the diagrams. Programmable switches was used to illustrate how the virtual devices and endpoint devices can be used to handle security threats. The use of Wired Equivalent Privacy (WEP) or WPA (Wi-fi Protected Access) which is a security algorithm for IEEE 802.11 wireless networks and other protocols for wired devices to target security challenges. As new ways to improve and achieve goals in a university is increasing, using modern technologies has made learning and teaching easier in which few universities have successfully adopted a standard networking model across Africa.

Keywords- Network Access Control (NAC), Switch, campus network, protocols

12. Paper 28021544: A Survey: Multimodal Systems of Finger vein and Iris (pp. 79-83)

Priyam Kaur Sandhu, CSE Department, PEC University of Technology, Chandigarh, India
Manvjeet Kaur CSE Department, PEC University of Technology, Chandigarh, India

Abstract — Biometric systems are the systems those enable automatic individual recognition which is based on behavioral or physical features belonging to a specific individual. All the biometric features have their limits to an extent and no biometric system is flawless. Therefore, the unimodal biometric systems have a lot of drawbacks. To solve the mentioned inconvenience and limitations and to enhance the level of security the multimodal biometric systems are employed. Personal identification process is a very important process that resides a large portion of daily usages. Human is a rich subject having many features that can be used for identification purpose such as finger vein, iris, and face etc. Finger vein recognition is an encouraging biometric recognition technique; the verification of individuals is done on the basis of the vein patterns present in the fingers. Iris recognition has moved under more focus due to its high reliability and efficiency in personal identification in past few years. This paper discusses the advantages of multimodal biometric system over unimodal biometric system along with the fusion of traits using different techniques. It also discusses the techniques employed on finger vein and iris. Finally, future scope is presented for the possible work that can be done in the field of finger vein and iris.

Keywords- finger vein; iris; PCA; multimodal; unimodal

13. Paper 30011503: Embedded Mobile Agent (EMA) for Distributed Information Retrieval (pp. 84-89)

Oguntunde B.O, Department of Computer Science, Redeemer's University, Ede, Osun State, Nigeria
Osofisan A.O, Department of Computer Science, University of Ibadan, Ibadan, Oyo State, Nigeria.
Aderounmu G.A, Department of Computer Science and Engineering, Obafemi Awolowo university, Ile-ife, Osun State, Nigeria

Abstract — Mobile agent paradigm has been recognised as a viable approach for building distributed applications. Mobile agents migrate through the network, execute asynchronously and autonomously, conserve bandwidth, achieve better load balancing, adapt dynamically to changes in their environment, are robust and fault tolerant. Existing agents run and execute on agent platforms also called, the Mobile Agent System (MAS), which provides run-time execution and support facilities for mobile agent to accomplish it tasks. These MASs from different vendors are different in language, design, and implementation and are not interoperable, this impedes the achievement of the full potentials of mobile agent paradigm. This work is aimed at providing a robust structure for

deploying mobile agents so they can execute independent of the MAS. We propose a lightweight agent to run in the kernel mode of the operating system as an operating system service, giving an impression of the agent directly communicating with the operating systems.

Keywords- embedded mobile agent, operating system service, lightweight agent, agent platform.

14. Paper 30011517: Fraudulent Electronic Transaction Detection Using Dynamic Kda Model (pp. 90-99)

Massoud Vadoodparast, Prof. Abdul Razak Hamdan, Dr. Hafiz
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia , 43600, UKM Bangi, Selangor, Malaysia

Abstract – Clustering analysis and Data mining methodologies were applied to the problem of identifying illegal and fraud transactions. The researchers independently developed model and software using data provided by a bank and using Rapidminer modeling tool. The research objectives are to propose dynamic model and mechanism to cover fraud detection system limitations. KDA model as proposed model can detect 68.75% of fraudulent transactions with online dynamic modeling and 81.25% in offline mode and the Fraud Detection System & Decision Support System. Software proposes a good supporting procedure to detect fraudulent transaction dynamically.

Keywords-component; Fraud detection, Data Mining, Clustering techniques, Decision Support System

15. Paper 28021512: Comparison between C++ console and graphic programming (pp. 100-107)

Zhengyu Lu, Department of Computer Science, Jilin University, China, Jilin Province, China
Dimitar Pilev, Department of Informatics, Chemical Technology and Metallurgy in Sofia, Sofia, Bulgaria

Abstract - Nowadays, the new students of Computer Science learn programming starting with C or C++ in the black-white console. At first, they may feel very proud of themselves by calculating or print the result on the Screen. As time goes by, they may feel tired of the black-white box and prefer something new. Then the problem comes, they should learn the basic C++ code by heart or just skip it to the graphic programming which sometimes seems more interesting than black-white console programming. And after the Graphic programming like MFC, CLR programming, we can even release the program and make it a small software that can be set up in their own laptop. It is really more interesting. So we will look at one example that we build the program through console and released-version graphic programming to see the difference and disadvantages and advantages of them.

Keywords-C++; Console programming; Graphic programming; Comparison;

A Novel Approach to Malware Detection using Static Classification

Sanjam Singla
Department of Computer Science
PEC University of Technology Chandigarh, India

Ekta Gandotra
Department of Computer Science
PEC University of Technology Chandigarh, India

Divya Bansal
Department of Computer Science
PEC University of Technology Chandigarh, India

Sanjeev Sofat
Department of Computer Science
PEC University of Technology Chandigarh, India

***Abstract*—Malware, commonly called computer virus, is one of the top security threats to the computer systems around the globe. These are evolving at a very rapid pace and are continually finding new ways to exploit and infect the systems of various enterprises and businesses. Malwares use different techniques to camouflage themselves to make their lifetime longer. In this paper, we present a simple technique based on static features extracted from Windows PE files. The features used are not only extracted from the header part of the malware but also from the payload i.e. body of malware. The static features used are a combination of Function Call Frequency and Opcode Frequency for differentiating malwares from clean files. This combination of features set makes it a new approach for malware detection which provides an accuracy of 97% for a dataset of 1,230 executables files including 800 malware and 430 cleanwares. For classification purpose, we use machine learning algorithms available in WEKA library. Based on the results obtained, we conclude that both features considered in this work play a significant role in distinguishing malicious files from clean ones.**

Keywords- Static Malware Analysis; Machine Learning; Classification;

I. INTRODUCTION

Malware is a hostile, intrusive or annoying software program designed to furtively gain access to a computer system without the owner's permission. According to McAfee 2014 Q3 threat report [1] there are over 307 new threats appearing every minute, and more than 5 are detected every second. According to the report, the total malwares broke the 300 million sample barrier in Q3 2014 i.e. the growth by 76% over the past year.

Nowadays, malwares are becoming more targeted and sophisticated. Advanced malwares are more dangerous as compared to traditional malwares as they are unknown, attack stealthily and are persistent in their attack [2]. Malware authors are often looking for one-time development of specific code to generate new variants of existing malware [3], instead of developing new malware from scratch as variants of existing malware can be developed easily and quickly.

Vulnerabilities in browsers and operating systems are often exploited by attackers. Also social engineering techniques are used to run malicious code on unsuspecting users. Malware authors often use obfuscation techniques [4] to evade detection by traditional defences like antivirus, firewalls, IDS, IPS, and gateways which uses signature based detection. Furthermore, different API calls used by malwares to perform malicious activities are also used by benign files to perform some operations and thus making it more difficult for the antivirus vendors to detect malwares accurately. In this paper, we extracted the features from both the header as well as payload of the malware which has not been used yet. We use Function Call Frequency and Opcode Frequency as static features to develop an automatic malware detection and classification system. A python script is used for extraction of mentioned features from the available data set, which is further classified into malwares and cleanwares with the help of machine learning algorithms available in WEKA. Our method gives an accuracy of above 97%.

Rest of the paper is structured as follows: In Section 2, we reiterate the literature in related area. Section 3 provides an overview about proposed malware classification system. Section 4 highlights the experimental set up along with the result analysis. Section 5 provides a brief discussion on results, followed by conclusion in Section 6.

II. RELATED WORK

This section describes the state of art on extracting various features for static analysis of malwares. These techniques were earlier used by developers in code optimization and reverse engineering but later used in malware analysis and detection.

Machine learning approaches like, Association classifiers, Support Vector Machines, Decision Trees, Random Forest, Naive Bayes and Clustering etc, are proposed for identifying and classifying new and unknown samples of malwares into known malware families.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thonnard *et al*. [5] developed a framework for detecting attack patterns in Honeynet data in order to find clusters of network traces sharing similar patterns within an attack data set, but limitation of his work is that the data collected through honeypots/ honeynets don't monitor network's normal traffic and were not able to cover the whole range of threats. Data mining concept for detecting malwares was first introduced by Schultz *et al*. [6]. Three different static features were used by them i.e. strings, byte sequence, Portable Executable (PE). In PE approach, features extracted were like, list of DLL function calls, and number of diverse system calls made within each DLL. Strings are taken out from executables based on static text based string matching which are encoded in program files. In byte sequence approach, sequence of n bytes extraction is done from an executable file. Later on Kolter *et al*. [7] improved the results by using overlapping byte sequence instead of non-overlapping byte sequence. With the help of this every possible sequence of bytes along with the frequency was extracted. In [8], [9], [10], [11], authors used byte n gram sequence to solve the malware classification problem. A new method for visualizing and classifying malware binaries as gray scale image using image processing techniques was proposed by Nataraj et al. [12]. A K-nearest neighbor method along with Euclidean distance method was used for malware classification. This method is fast but has limitation. As this method uses global image based feature, an attacker can implement countermeasures to thrash the system. For automated malware classification, Kong *et al*. [13] presented another framework which was based on structural information of malwares. After extracting the features based on function call graph, similarity is calculated for the two programs and discriminate distance metric learning was used which groups the malware samples belonging to alike family. In [14], [15], the authors analyzed the sequence of API calls used in PE code. Tian *et al*. [16] classified Trojan based on function length frequency, where length of function is the measure of the number of bytes in the code. Siddiqui *et al*. [17] classified worms via variable length instruction chain along with machine learning technique. In [18],[19], proposed methods for analyzing entropy of PE sections in an executable. Saini et al. [20] used suspicious section count and function call frequency as the features to distinguish malwares from clean ones. They have used machine learning algorithms as available in WEKA library for classification purpose. They achieved an accuracy of more than 98%.

In our work, we used four classification algorithms for distinguishing malwares from cleanwares: Decision Table (DT), J48 tree, KStar, MultilayerPerceptron (MLP), from WEKA (Waikato Environment for Knowledge Analysis) [21]. These algorithms have a wide range of techniques available for classification.

III. Malware Analysis And Classification System

In this section, we present our approach to the classification problem using machine algorithm available in WEKA, over a data set of 1,230 executables after extracting the static features as mentioned in following sections.

A. The Data Set

We used a sum of 1,230 executable files including 800 malware and 430 clean files. The malware samples are provided by University of California, Santa Barbara, (http://seclab.cs.ucsb.edu/) on request and rest are collected from available online data sources like http://www.nothink.org/honeypots/malware-archives. The clean files are collected from the System32 directories of Windows 2000, Windows 2003, Windows XP, Windows 7 and Windows 8.

B. Static Features Used

This section explains the features which we used for distinguishing malwares from cleanwares. We not only used the extracted features from the header of the malwares but also from the payload i.e. body of the malwares. Function Call Frequency and Opcode Frequency features are extracted from each executable including malware and clean ware using a python script, which are then passed to automatic classification system for further analysis.

C. Function Call Frequency

Malware authors use obfuscation techniques to evade detection by Antivirus vendors. Obfuscation techniques used like dead code insertion, instruction replacement, register substitution, code transportation, and instruction permutation. Normally legitimate programs include a number of strings but obfuscated programs contain very few strings. If a particular searched file is packed or obfuscated then the number of strings will be less, thus suggesting that it may be malicious. Earlier, in [20], [23] authors have used function call frequency as one of the features to distinguish malwares from clean ones, we also in this work used the same feature but with a different combination for distinguishing the data.

In our approach, we analyzed the number of function calls made by malware and clean samples. We first created a global list of all known function calls and then match all the readable strings extracted from the analyzed files, which gives the exact numbers of function calls made by that executable. For instance, global list is considered to be made up of {'CheckRemoteDebuggerPresent', 'IsDebuggerPresent', 'LoadLibrary', 'WriteFile', 'CreateProcess'} functions based on which classification is done.

D. Opcode Frequency

Opcode basically means operational codes that are used in the files in order to perform some operations. In order to detect the advanced malwares like polymorphic and metamorphic, the opcode frequency can be used in analysis for their detection.

In our approach, we analyzed the number of opcode calls made by malware and clean samples. We first created a global list of all opcodes that are not common in benign files and then match all the readable strings extracted from the analyzed files, which gives the exact numbers of opcode calls made by that executable. Due to uncommon opcode list it is easy to distinguish between clean and malicious file as frequency of these opcodes is high in malwares as compared to that in clean files. For instance, the global list i.e. uncommon is made up of {'int', 'nop', 'rdtsc', 'sbb', 'shld', 'fdvip', 'imul', 'pushf',

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

'setb', 'fild'}. However, some common opcodes like {'mov', 'pop', 'call', 'test', 'jmp', 'lea'} are also considered for achieving better accuracy in the classification process.

IV. Experimental Setup and Result Analysis

This section describes complete set-up for malware analysis and classification. It also includes the experimental result analysis. The experiment is performed on local system with configuration having Ubuntu 14.04 as operating system with 16GB RAM and 1TB hard disk. After extracting the static features from the data set, we used four classifiers: Decision Table and J48 tree, KStar, MLP available in WEKA library for differentiating malwares from cleanwares as mentioned in section 2. WEKA is a set of machine learning programs implemented in Java. It is developed at the University of Waikato and is freely available under the GNU General Public License. The standard input format accepted by WEKA is ARFF (Attribute-Relation File Format). We used WEKA 3.7 for Windows OS version.

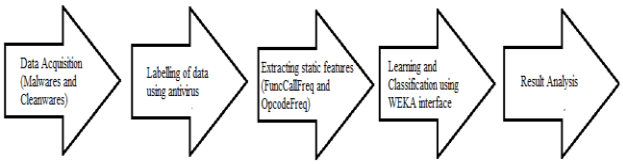


Figure 1. Malware analysis and classification system

Figure 1 illustrates the automated detection and classification system. We collected a large set of executable programs from a wide range of sources and separated them into malicious and benign executables using antivirus. For each executable file, including both malware and cleanware, we extracted Function call frequency and opcode Frequency as features from the complete data set using a Python script. These features are then used for differentiating malwares from cleanwares using WEKA [21] interface as mentioned in section 2. In our experiments, we used 10-fold cross validation for estimating our results as it is best method for estimation for likely predictions over unseen data in Data Mining. In 10-fold cross validation, the novel sample is arbitrarily partitioned into 10 sub samples each having equal size. Out of 10 sub samples, 9 are used for training and 1 is retained as validation data for testing purpose. The process is then repeated 10 times with each of the 10 sub samples used precisely once as the validation data. The average of these results is calculated in order to achieve a good measure of how the algorithm performs over the entire data set.

A. Experimental Results

The classification is done using machine learning algorithm available in WEKA. Detection rate is measured using three parameters: False Positive (FP), False Negative (FN) and the Detection Rate Accuracy.

Table 1 shows the weighted average of the experimental results for the Function Call Frequency (FunCallFreq) feature.

Table 1. Classification results using FuncCallFreq

Classifier	**Classification Results**		
	FP	***FN***	***Accuracy (%)***
MLP	0.294	0.025	88.14
Kstar	0.068	0.03	91.56
J48	0.021	0.031	96.77
DT	0.014	0.031	97.37

MLP classifier shows the worst accuracy whereas DT shows the highest accuracy rate of 97.37% followed by J48 which gives 96.77% accuracy.

Table 2 shows the weighted average of the experimental results for the Opcode Frequency (OpcodeFreq) feature.

Table 2. Classification results using OpcodeFreq

Classifier	**Classification Results**		
	FP	***FN***	***Accuracy (%)***
MLP	0.250	0.098	84.71
Kstar	0.247	0.092	85.33
J48	0.198	0.097	86.72
DT	0.205	0.101	86.26

MLP classifier shows the worst accuracy whereas J48 shows the highest accuracy rate of 86.72% followed by DT which gives 86.26% accuracy. To further improve our results, we combined both the features and tested them on different classifiers.

Table 3 shows the weighted average of the experimental results for both the features taken together.

Table 3. Classification results using both FuncCallFreq and OpcodeFreq

Classifier	**Classification Results**		
	FP	***FN***	***Accuracy (%)***
MLP	0.275	0.026	88.69
KStar	0.156	0.026	92.84
J48	0.033	0.027	97.07
DT	0.014	0.031	97.47

With both the features combined together, DT gives the best accuracy i.e. 97.47 % followed by J48 which provides an accuracy of 97.07%. After comparing table 3 with table 1 and 2, we see that on an average the number of both FP and FN are reduced when combined features are considered.

Figure 2 shows the comparison of classifiers with respect to the features considered in our experiment.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

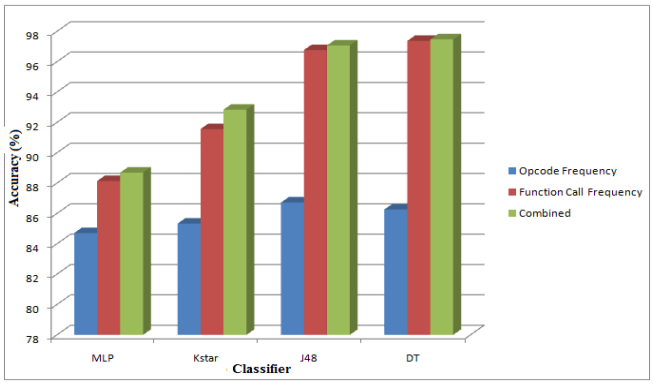


Figure 2. Comparison of accuracy of different classfiers using FuncCallFreq and OpcodeFreq

V. CONCLUSION

In this paper, the function call frequency and opcode frequency are used as static features for distinguishing malwares from cleanwares. The results obtained for our experiments indicate that the function call frequency is more robust feature than opcode frequency. Further, when we combine both features, DT proves to be the best machine learning technique equipped to classify our dataset. Though we get good results with static analysis, yet we understand that our technique may not be effective to classify zero day malwares and furthermore disassembling for extracting opcodes is a time consuming task. In future, we would come up with an optimized dynamic analysis technique, which will be able to classify advanced and stealth malwares.

REFERENCES

[1] Mcafee labs threats report:
http://www.mcafee.com/in/resources/reports/rp-quarterlythreat-q3 2014.pdf

[2] E. Gandotra, D. Bansal and S. Sofat, "Malware anlaysis and classification: A survey," Journal of Information Security, Vol. 5, No. 2, pp. 56-64, April 2014.

[3] S.Treadwell and M. Zhou, "A Heuristic Approach for Detection of Obfuscated Malware," in Proceedings of the 3rd International Conference on Intelligence and Security Informatics, IEEE, pp. 291–299, 2009

[4] K. Yim, "Malware Obfuscation techniques: A brief survey," In Proceedings of International Conference on Broadband Wireless Computing Communication and Applications (BWCCA), Fukuoka, pp. 297-300, November 2010.

[5] Thonnard and M. Dacier, "A framework for attack patterns' discovery in honeynet data," Digital Investigation: The international Journal of Digital Forensics and Incident Response, vol. 5, pp. 128-139, September 2008.

[6] M. Schultz, M. Eskin, E. Zadok, and F. Stolfo, "Data Mining Methods for Detection of New Malicious Executables," In Proceedings of 2001 IEEE Symposium on Security and Privacy, IEEE, Oakland, CA, pp. 38-49, May 2001.

[7] J. Kolter, M. Maloof, "Learning to Detect Malicious Executables in the Wild," In Proceedings of 10th ACM International Conf. on Knowledge Discovery and Data Mining (SIGKDD), ACM New York, NY, USA, pp. 470– 478, 2004.

[8] R. Moskovitch, D. Stopel,, C. Feher, N. Nissim, and Y. Elovici, "Unknown Malcode Detection via Text Categorization and the Imbalance Problem," In Proceedings of 6th IEEE International Conference on Intelligence and Security Informatics (ISI), Taiwan, pp. 156–161, 2008.

[9] Santos, Y. Penya, J. Devesa and P. Bringas, "N Grams-Based File Signatures for Malware Detection," In Proceedings of 11th International Conference on Enterprise Information Systems (ICEIS), AIDSS, Milan, Italy, pp. 317–320, 2009.

[10] Y. Zhou and W. Inge, "Malware Detection using Adaptive Data Compression," In Proceedings of 1st ACM workshop on AISec, ACM New York, NY, USA, pp. 53– 60, 2008.

[11] Y. Ye, Y. Li, Y. Chen and Q. Jiang, "Automatic Malware Categorization using Cluster Ensemble," In Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July, pp. 25-28,95-104, 2010.

[12] L. Nataraj, S. Karthikeyan, G. Jacob and B. Manjunath, "Malware Images: Visualization and Automatic Classification," In Proceedings of 8th International Symposium on Visualization for Cyber Security, ACM New York, USA, 2011.

[13] D. Kong and G. Yan, "Discriminant malware distance learning on structural information for automated malware classification," Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems, ACM New York, USA, pp. 347-348, 2013.

[14] C. Wang, J. Pang, R. Zhao and X. Liu, "Using API Sequence and Bayes Algorithm to Detect Suspicious Behavior," In Proceedings of International Conference on Communication Software and Networks, IEEE Computer Society Washington, DC, USA, pp. 544–548, 2009.

[15] J. Xu, A. Sung, P. Chavez and S. Mukkamala, "Polymorphic Malicious Executable Scanner by API Sequence Analysis," In Proceedings of 4th International Conference on Hybrid Intelligent Systems, IEEE Computer Society Washington, DC, USA, pp. 378-383, 2004.

[16] R. Tian, L .Batten, R. Islam and S. Versteeg, "An automated classification system based on the strings of Trojan and virus families," In Proceedings of 4th International Conference on Malicious and Unwanted Software, Montréal, Quebec, Canada, pp. 23-30, 2009.

[17] M. Siddiqui, M.C. Wang and J. Lee, "Detecting Internet Worms Using Data Mining Techniques," Journal of Systemics, Cybernetics and Informatics, Vol. 6, pp. 48-53, 2009.

[18] R. Lyda and J. Hamrock, "Using Entropy Analysis to Find Encrypted and Packed Malware," IEEE Security & Privacy 5, pp. 40-45, 2007.

[19] X. Pedrero, I. Santos, B. Sanz, C. Laorden and P. Bringas, "Countering Entropy Measure Attacks on Packed Software Detection," In Proceedings of 9th IEEE Consumer Communications and Net-working Conference (CCNC2012), Las Vegas, NV, pp. 164-168, 2012.

[20] Saini, E. Gandotra, D. Bansal and S. Sofat, "Classification of PE files using static analysis" SIN'14, Glasgow, Scotland, UK, ACM 2014.

[21] S. Michael and H. Andrew, "Basic Static Techniques, 'Practical Malware Analysis' No Starch Press, San Francisco, pp. 21-24, 2012.

[22] Avira antivirus:
http://www.avira.com/en/avira-free-antivirus

[23] E. Gandotra, D. Bansal and S. Sofat "Integrated Framework for Classification of Malwares," SIN'14, Glasgow, Scotland, UK, ACM 2014.

AUTHORS PROFILE

Sanjam Singla is a master student in CSE department of PEC University of Technology, Chandigarh. He completed his graduation from RBIEBT, Mohali Campus under Punjab Technical University, Jalandhar in year 2013.

Ekta Gandotra is a doctoral student in CSE department of PEC University of Technology, Chandigarh. She completed her M. Tech. in Computer Science and Engineering from Kurukshetra University, Kurushetra in year 2008. She is pursuing her Ph.D. in the area of Cyber Security.

Divya Bansal is Associate Professor with PEC University of Technology, Chandigarh. She is also the Associate Coordinator for Cyber security Research Centre, Chandigarh where she leads a research and development team working on wireless security. She did her Masters in Engineering in

Computer Science & Engineering from PEC University of Technology and also her Ph.D. in the area of security in wireless networks from PEC. Her research interests include areas in Information Security such as Cyber Crime & Investigations, Cyber Warfare and Cryptology, Cloud Computing, Wireless Networks. She has done several government sponsored research projects under the capacity of Principal Investigator. She has over 50 research publications in reputed journals and conference proceedings.

Sanjeev Sofat is a Distinguished Professor with PEC University of Technology, Chandigarh. He is also the Coordinator for Cyber Security Research Centre, Chandigarh. He completed his M.E in Computer Science from ISM, Dhanbad in the year 1993. He received his Ph. D. degree from Kurukshetra University in 2005. His current areas of research are Computer Networks & Information Security. He is a member of IEEE and CSI.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Performance Evaluation of IEEE 802.15.6 Improvised and Scheduled Access Modes for Remote Patient Monitoring Applications

Anas Bouayad, Nour El Houda Chaoui, Moulhime El Bekkali, Mohammed El Ghazi
Transmission and Treatment of Information Laboratory
USMBA, FST
FEZ, MOROCCO

***Abstract*- In the past few years, substantial improvement has been made in the medical field to integrate communication and information technology especially Wireless Body Area Networks (WBANs) in healthcare systems for remote patient monitoring (RPM). The wide diffusion of healthcare monitoring systems allows continuous patient to be remotely monitored and diagnosed by doctors. WBANs have shown great potential in improving healthcare quality, and thus have found a wide range of applications from ubiquitous health monitoring and computer assisted rehabilitation to emergency medical response systems. WBAN operates in close vicinity to, on, or inside a human body and supports a variety of medical applications. Different standards and communications protocols are used in WBANs such as IEEE 802.15.6. The IEEE 802.15.6 standard offers a flexible superframe structure that can be adjusted by the hub to suit the communication requirements of the network and applications. However, the standard leaves the higher level questions open such as: should we use contention-based, scheduled, or improvised access, and under what conditions should we use them. To exploit these access modes we should have a clear understanding of their parameters and operating characteristics.**

In this work, we are interested in studying access methods and the polling access mechanism used in MAC layer of the IEEE 802.15.6 standard and the proposition of suitable access methods and parameters should be used to increase the performance of the MAC protocol in terms of successful received packets and low latency. Performance evaluation will be based on the simulation of a short range wireless Body Area Network based solution implementing the IEEE 802.15.6. Simulation will be performed on OMNet++ with the Castalia simulator.

Keywords: RPM, wireless Body Area Networks, IEEE 802.15.6, medium access control (MAC) protocols, access methods, polling.

I. INTRODUCTION

There are tens of thousands of remote areas in developing countries, inter alia, Morocco, where the availability and exchange of data related to health may contribute to the prevention of the disease and save the life of thousands of people.

Appropriate monitoring of environment variables is also necessary to implement preventive measures at the local level or through appropriate government policies. The areas of health and food safety are raised on the agenda of objectives Millennium development Goals (MDGs) of the UN (United Nations) [1].

There are a variety of possible technical tools contributing to the solution of a social problem that is in a particular context. The questions of the complexity of the equipment, finance, local expertise, infrastructure deployment, were obstacles to make the most of these durable solutions.

The goal of our research is to find innovative and sustainable solutions to help improve the quality of health services in developing countries where the lack of qualified and competent staff (nurses, doctors, ...) and medical equipment are real problems facing developing countries, especially in the villages far from major health centers.

A wireless sensor network (WSN) is a communication network composed of wireless sensor devices. These devices essentially are low cost, low power, multi-functional, small sized and communicate over short distances. Typically these

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

devices serve as nodes in a wireless network and are deployed randomly in a given area. Nodes establish connectivity with each other dynamically after deployment and do not follow a pre-determined topology and a specified protocol of communication. Therefore WSN are self-organizing in nature and are suitable for many fields and areas. One application of WSN is the monitoring of the health of patients remotely, Known as WBAN (Wireless Body area Network). Wireless Body Area Networks (WBANs) have emerged as a solution better suited for biological signal monitoring. They allow for mobility, usability, and comfort for the users. Furthermore, patients do not need to stay in hospital to be monitored, which reduces health costs. These benefits have motivated the growth of several WBAN applications in medical field. A significant amount of recent research has been done in the field of wireless body area networks with many researchers who propose different types of solutions for patient supervision. [2, 3, 4, 5, 6] are examples of such systems.

Within this context, the objective of this work is to model and simulate a heterogeneous wireless sensor network allowing the measurement and the transmission of short-range data collected by the environmental sensors. The planned network will be deployed in a real world setup, precisely, in remote hospital centers and transmit health data and alert messages caused by a malfunction of physiological parameters to central hospital center via a continuous monitoring. So a limited scale, up to 20 nodes, seems to be sufficient for this monitoring application. These nodes exchange data between them according to a communication protocol that optimizes energy consumption, transmission delay and loss of information. Another principle to consider is that when any node fails, the network should repair automatically and must run normally with a minimal loss of information.

In fact, from the network point of view, key emphasis of this work is on WBANs which tries to provide low power, low cost and short-range solutions. Among them, IEEE 802.15.6is considered as a promising way in terms of energy saving and guaranteed medium access. Therefore, we consider IEEE 802.15.6 as a starting point for our work. We optimized the parameters of the IEEE 802.15.6 physical layer and medium access control layer to better suite our specific constraints. The MAC layer has a fundamental and significant impact in a protocol stack. The upper layers including network layer, transport layer, application layer, etc. will be considered after a robust MAC layer.

The paper is organized as follows: related works are discussed in section II, section III details our proposed system architecture. Section IV highlights the IEEE 802.15.6 MAC standard. Section V presents simulation parameters. Section VI focuses on performance evaluation of the configured access modes and parameters. Finally, the summary of the analysis, conclusion and the future works are given in section VII.

II. RELATED WORKS

There are already several prototypes of WBANs for remote health monitoring. For example:

CareNet project [7] an integrated wireless sensor environment for remote healthcare that uses a two-tier wireless network and an extensible software platform. CareNet provides both highly reliable and privacy-aware patient data collection, transmission and access.

Ayushman project is a sensor network based health monitoring infrastructure [8]. Ayushman provides a medical monitoring system that is dependable, energy-efficient, secure, and collects real-time health data in diverse scenarios, from home based monitoring to disaster relief.

The Medical Emergency Detection in Sensor Networks (MEDiSN) project [9] utilizes a wireless sensor network composed of a network gateway, physiological monitors (PMs), and relay points (RPs), to monitor the health and transmit physiological data of patients. The PMs are sensor devices which collect, encrypt and sign patients' physiological data (e.g., blood oxygen level, pulse, ECG, etc.) before transmitting them to a network of relay points that eventually forwards the data to the network gateway.

The European Community's MobiHealth System demonstrated the Body Area Network (BAN) consisting of software programs, hardware devices (including sensors) and Bluetooth communication between devices such as the MobiHealth GPRS Pregnancy Body Area Network [10]. The challenges of wireless networking of human embedded smart sensor arrays for a proposed retina prosthesis are described in [11].

However, studies on the use of IEEE 802.15.6 for remote health monitoring still few, and existing solutions need to be reviewed for more optimizations.

For example, the work done by Timmons and Scanlon which propose a BAN MAC, while at the same time arguing the non-suitability of the 802.15.4 MAC for BAN [12].

A new on-going project called BANET [13], which has as major objectives to provide a framework for Body Area Networks, define a reliable communication protocol, optimize BAN technologies and enhance energy efficiency of network

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

components. The Project is led by CEA-Leti. It aims at defining precise frameworks to design optimized and miniaturized wireless communication systems. These body area networks target the medical field.

In addition, performance analysis of MAC access methods still few compared to other standard. For example: An analytical model was developed for performance evaluation of the IEEE 802.15.6 standard [14]. This analysis shows that adopting appropriate user priorities and Exclusive and Random Access Phases (EAP1and RAP1) lengths enhance the performance of the network. In [15] we find a study that investigates the impacts of channel fading and diversity in the MAC layer of an IEEE 802.15.6 CSMA/CA based WBAN.

III. SOLUTION ARCHITECTURE

In this section, we describe our architecture solution which enables a healthcare institution, such as a Central Hospital Center (CHC), to manage data collected by WSN for sick patient supervision in Remote Healthcare Centers (RHC).

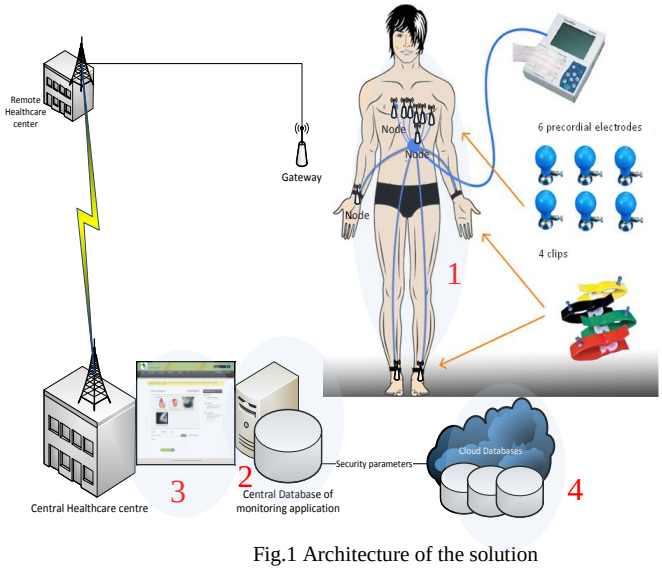


Fig.1 Architecture of the solution

The solution aims to store a very large amount of data generated by sensors in the cloud. As these data are very sensitive, a new security mechanism to guarantee data confidentiality, data integrity and fine grained access control should be defined.

In the architecture described in Fig.1, we consider two categories of users, healthcare professionals and patients, and is composed of the following components: (1) the WBAN system which collects health information from patients, (2) the monitoring applications which allow healthcare professionals to access to stored data, (3) the Healthcare Authority (HA) which specifies and enforces the security policies of the healthcare institution and (4) the cloud servers which ensure data storage. By storing data on the cloud, our architecture offers virtually infinite storage capacity and high scalability.

IV. IEEE 802.15.6 MAC standard

The IEEE 802.15.6 [16] is a standard for Body Area Network, which operate in and around the human body (but not limited to humans). According to the IEEE, the new standard is more flexible and can be used for both medical and non-medical applications. It promises a maximum throughput of 10 Mbit/s, combines safety, reliability, quality of service, low consumption and protection against interference which render it suitable to satisfy multiple applications of personal wireless networks (WBAN, Wireless Body Area Networks). It covers the physical (PHY) and Medium Access Control (MAC) layers. The first one offers operation modes such as narrowband (NB), ultra-wideband (UWB), and human body communication (HBC). The second one offers different operation modes and medium access methods. In this section, we try to present the medium access protocol described in the standard which specifies a medium access with the different access modes and their access phases and access mechanisms, and we will focus on the polling mechanism.

A. Access modes and mechanisms

A hub may operate in three different modes as described below:

- Beacon mode with beacon period (superframe) boundaries; at the beginning of every superframe a beacon is transmitted on the medium to provide time referenced allocations. Each superframe is divided into access phases (APs) as illustrated in Fig. 2. A superframe includes exclusive AP 1 (EAP1), random AP 1 (RAP1), type-I/II AP, exclusive AP 2 (EAP2), random AP 2 (RAP2), type-I/II AP, and contention AP (CAP). Each access phase, except RAP1, may have a zero length.
- Non-beacon mode with superframe boundaries in which the hub may have only the type-I/II access phase.
- Non-beacon mode without superframe boundaries in which the hub only provides unscheduled polled allocations.

Medium access mechanisms of the IEEE 802.15.6 standard can be divided into four categories; random

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

access (connectionless contention-based access), improvised and unscheduled access (connection less contention-free access), scheduled access and variants (connection-oriented contention free access).

1) Random access (connectionless contention-based access)

In EAP1, RAP1, EAP2, RAP2, and CAP, allocations may only be contended allocations, which are non-reoccurring time intervals valid per instance of access. The EAPs are reserved for emergency high priority traffic while the RAPs are used for nonrecurring transfers.

The access method for obtaining the contended allocations shall be:

- CSMA/CA if pRandonAccess is set to CSMA/CA.
- slotted Aloha access if pRandonAccess is set to Slotted Aloha

2) Improvised and unscheduled access

a) Unscheduled access

A hub may employ unscheduled polling and posting access to send polls or posts at any time to grant polled or posted allocations either in beacon mode or non-beacon mode, so long as the addressed nodes indicated that they will always be in active state through their last transmitted MA Capability field (i.e., with Always Active bit set to 1).

A node that has so indicated shall constantly be in active state ready to receive unscheduled polls or posts.

b) Improvised access

A hub may employ improvised polling and posting access to send polls or posts at previously announced times based on predefined Table to grant polled or posted allocations, either in beacon mode or non-beacon mode for on-demand contention-free frame exchanges outside the scheduled allocations within their body area network (BAN).

A polled or posted allocation contains an explicit or implicit time interval that does not reoccur subsequently without the hub invoking another instance of improvised access.

Unscheduled and improvised transfers occur in the type I/II access phases.

3) Scheduled access and variants

A node and a hub may employ scheduled access to obtain scheduled uplink allocations and scheduled downlink allocations, scheduled-polling access to obtain scheduled bilink allocations and polled allocations therein, and delayed polling access to obtain delayed bilink allocations and polled allocations therein.

The allocations may be: 1-periodic or m-periodic allocations, but a node shall not have both 1-periodic and m-periodic allocations in the same body area network (BAN).

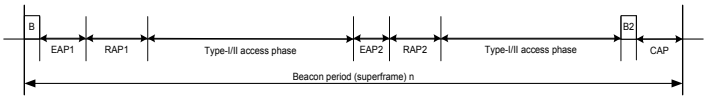


Fig. 2 Layout of access phases in a beacon period (superframe) for beacon mode

Scheduled transfers occur in the type I/II access phases.

B. Polling mechanism

Polling is a media access method that is used in many types of wireless networks.

Polling resembles a well-ordered meeting in which the chairman must recognize an attendee before that person is allowed to speak. The chairman's responsibility is to maintain order in the meeting and ensure that each person who wants to speak has an opportunity to do so. Polling is most closely associated with point-to-point wireless networks. By using polling, one device is designated as the primary device (coordinator). Primaries also are known as the channel access administrators, controllers, or masters. All access to the network is controlled by the coordinator.

In the IEEE 802.15.6 standard, the time is divided into beacons period. When the hub start to construct the network he sends a management frame, to all nodes, which handles all information about the BAN, such as BAN ID, number of time slots in a beacon, duration of each time slot, the length of each access phase (EAP, RAP, CAP), etc. This BAN information serves the nodes to choose the access technique to be used in different phases. A hub may send polls and grant type-I or type-II polled allocations to a node only if both of them support polling access of the corresponding type as indicated in their last exchanged MAC Capability field. So that a node can get polled allocation, it should set in the frame it is transmitting, the More Data field to value one. The node should also set the Ack Policy field to I-Ack or B-Ack in some management or data type frames

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

being transmitted. This enables the hub to send the node an immediate or future poll at an announced time through an I-Ack+Poll or B-Ack+Poll frame. To grant an immediate polled allocation to a node, a hub shall send to the node a Poll or T-Poll frame when appropriate or an I-Ack+Poll or B-Ack+Poll frame when required to return an acknowledgment. The process is presented in the Fig 3.

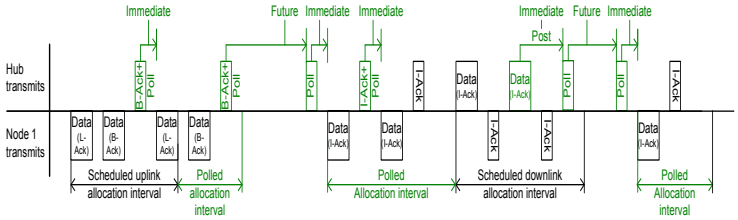


Fig. 3 Example of polled allocation

C. Sleep mode

Nodes in the network sleep most of their lifetime. They wake up only to transmit Data. As soon as nodes finish transmitting packets, they start sleeping again. The time at which a node wakes up is determined by the hub. The hub sends a poll packet to a node according to the poll schedule stored in the hub. Ideally, a node need wake up just at the moment it should receive the poll packet from the hub. If the node wakes up earlier, it will have to stay awake to receive the poll packet from the hub causing unwanted energy losses. If the node wakes up after the poll packet is sent by the hub, the poll packet will be lost and the polling mechanism fails. The hub has to ensure that the node receives the poll packet. The hub therefore sets a sleeping time for each node after the transmission of the packets. The node should sleep for the time specified by the hub after which it wakes up at the right moment to receive the poll packet. However, due to variations in times for which packets are transmitted and because of clock synchronization problems, the node may wake up before or after the stipulated time for sending the poll packet by the hub.
A mechanism is developed in [17] whereby a sleeping node wakes up at the right moment to receive poll packet.

V. SIMULATION PARAMETERS

The simulation framework we choose is the Castalia open source simulator [18]. All simulations described in this paper are released with Castalia 3.2, assisting with the reproducibility of the results. Fig.4 shows the simulated network topology used throughout our simulations. One coordinator node at the right of the human body, and ten sensor nodes sending packets of 128bytes (including overhead) to the coordinator.

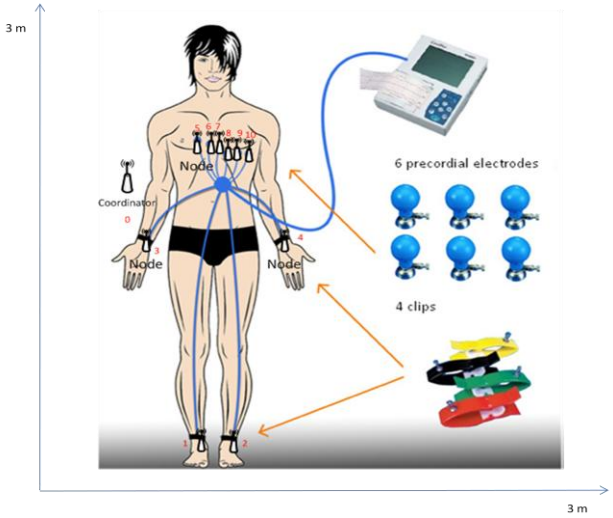


Fig.4 Simulated Network Topology

The radio parameters we define that meet with the IEEE 802.15.6 radio proposal [19] are: frequency, data rate, modulation type, bits per symbol, bandwidth, noise bandwidth, noise floor, sensitivity and power consumed. We also define Tx levels in dBm and mv, delay transition between states, power transitions between states, and sleep levels. Table 1gives the various radios parameters defined.

Table 1
Radio parameters defined

Data rate	1,024Kbps
Modulation	Diff QPSK
Rx sensitivity	−87dBm
Noise bandwidth	1MHz
Noise floor	−104dBm
Tx power	−10dBm
CCA time	1ms
Tx→Rx and Rx→Tx(transition times)	20µs
Rx→Sleep, Tx→Sleep (transition times)	0.194ms
Sleep→Rx, Sleep→Tx (transition times)	0.05ms
Tx (power consumed) Rx (power consumed)	3mW 3.1mW
Tx→Rx, Rx→Tx (power consumed)	3mW
Sleep→Rx, Sleep→Tx Rx→Sleep, TX→Sleep (power consumed)	1.5mW
Sleep power level	0.05mW

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The effect of path loss is considered from [20]. The channel temporal variation is considered with the existing model of Castalia.
For the MAC Layer, We have also implemented most aspects of the 802.15.6 MAC standard described in the "MAC and Security Baseline Proposal", IEEE 802.15 Documents [21].

Table 2
MAC default parameters defined

Slot allocation length	10 ms
Allocations slots in a beacon period	32 slots
Requesting slots per node	3 slots
Total allocation slots	30
Contention based access slots	2
Buffer MAC	48 packets
Retransmission packets tries	2
Polling	Enabled

Table 2 give the most important default parameters used in the simulation scenario.

The 128byte data packet needs 1ms to be transmitted with the BAN radio used. Since though all data packets are required to be acknowledged, the total time for a packet (TX + ACK + radio state transition times) is 1.16ms. This means that there are 8 packets fitting in each allocation slot.

If each node gets allocated 3 slots and gets to use 1/10 of the 2 remaining slots then each node could get 3.2 time slots in each beacon period. As a result each node could transmit 8 packet/slot * 3.2 slots = 26.21 packets (1). The beacon period duration is equal to 320 ms, so the number of beacon per second is equal to 3.125. (2)

The Packet data rate can be calculated from (1) & (2): $R_{packet} = N_{packets\ sent/beacon} * N_{beacon/secd}$
= 80 packets /sec

All runs last 51 sec (50sec for data and 1sec used for network setup). Each of the cases was executed 10 times with different random seeds.

VI. SIMULATION RESULTS AND PERFORMANCE EVALUATION

We observed performance metrics such as received packets, Data packet breakdown and latency. The results presented are averages on all nodes. Although, in this paper, we do not have the space to present the differences between nodes. (as a result of the different link qualities)

A. *Received packets*

The received packets graph presented bellow (Fig.5) shows the average packets received per node (only node 0 receives packets but it receives them from multiple nodes) for different length of scheduled access slots and random access slots. Nodes are sending packets for 50 sec so if we had perfect reception we would reach 4000 packets per node for the 80packets/sec/node case.

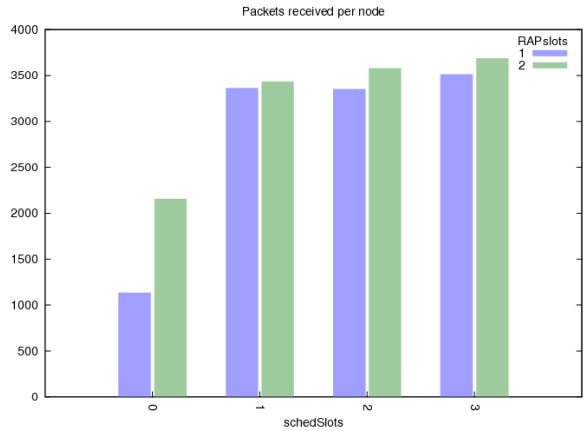


Fig.5(a) Packets received per node(polling enabled)

We notice a received packets efficiency of 92% with the case of three scheduled access slots and two random access slots. As shown in the graph, better performance (packets received) is obtained when the number of scheduled access increases.

Also we notice that the protocol performs better when the polling mechanism is turned on. This is something to be expected as the polling mechanism makes a more efficient use of the wireless channel and is reducing interference. We can see (Fig.5 (b)) the big difference for the case of scheduled access slot = 0.

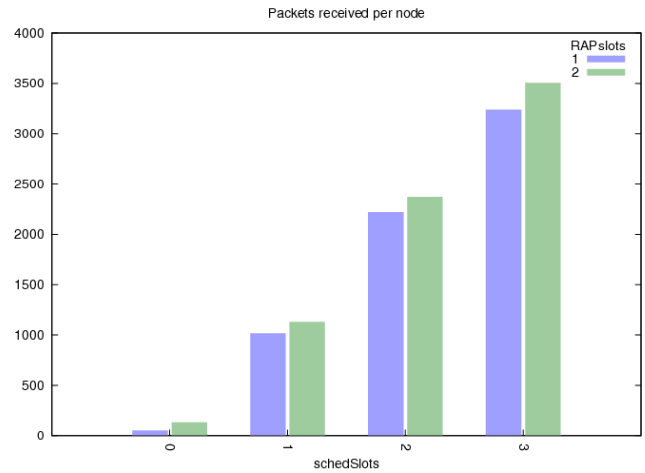


Fig.5(b) Packets received per node (polling disabled)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Latency

The latency of a successful packet is an important performance metric. If a packet is successful but arrives beyond a certain latency threshold the packet might be useless. The latency histogram presented bellow shows the distribution of latency (Fig.6) for the rate of 80packets/sec/node and different length of scheduled access slot (0, 1, 2, 3).

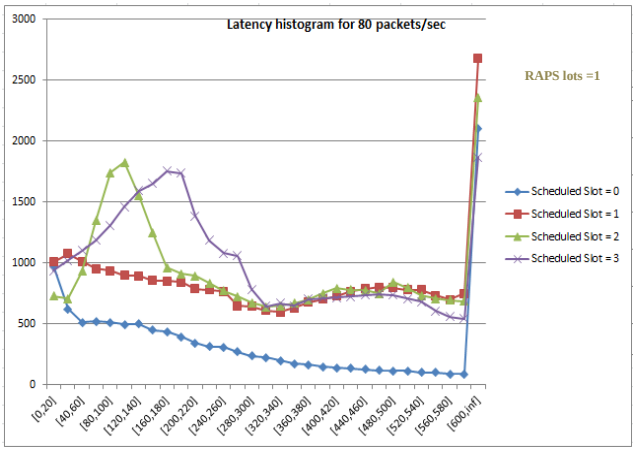


Fig.6 (a): latency distribution for different scheduled access length (RAP slots = 1).

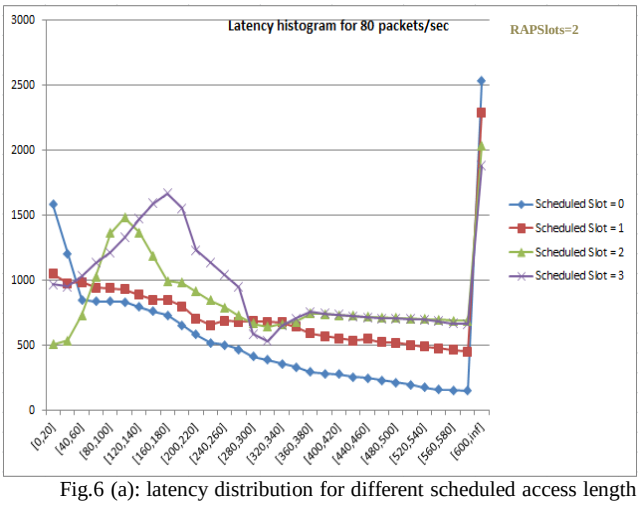


Fig.6 (a): latency distribution for different scheduled access length (RAP slots = 2)

In the graphs above (Fig.6), we observe high performance when scheduled access length is equal to 3. We could see that we have around 75% of packets that are delivered within 300 ms for both cases (RAPslots=1 and RAPslots=2). This is because polling method does not affect the scheduled access period, it provides good time synchronization. When the pool interval arrives, the hub gives the permission to the node to starts transmission by sending the poll frame. Before the node receives the poll and starts data transmission, the hub has already finished the transition from TX mode to RX mode.

We notice also that for longer period of polling, the packets are delivered with high latency.

C. Packets Breakdown

We observe the packet breakdown at the MAC layer of the senders. The packets are divided into five categories: 1) success on the first time(i.e., an ack was received on the first transmission attempt), 2)success (i.e., an ack was received on the second or more transmission attempt), 3) No Ack (i.e. packet received no ack and was transmitted at least once on the radio), 4) Channel busy (i.e., packet failed because the CSMA mechanism never found the channel free, in all transmission attempts), 5) Overflow (i.e., the MAC buffer was full so the packet was rejected).

Fig.7 presents results for different cases in a comprehensive manner.

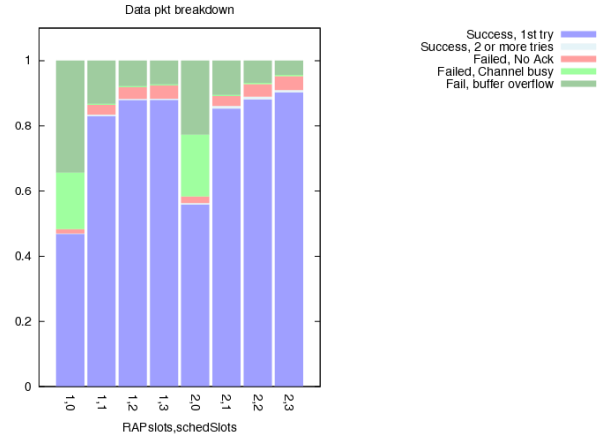


Fig.7: Data packets breakdown at the MAC layer

The horizontal axis of each graph varies the length of scheduled access slots from 0 to 3, and random access slots from 1 to 2. The vertical access shows the data packets breakdown. The results are shown as fractions of 1.

The first characteristic we notice is the increase of successful packets rate and the decrease of buffer overflow as the duration of scheduled access increase. This is due to that polling mechanism relies on control messages (poll and future poll frame). Future polls are transmitted multiple times with the ACKs at the last slot a node has transmit access, while a poll message is transmitted only once at the beginning of the polled access. So it is more vulnerable, because in the case of the failure of the poll frame, the allocated slot is wasted and no node

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

could use it. Also if a beacon is lost we don't have any activity for that node on that frame and arriving packets are just buffered. Form the results of our simulation we found that around 84 % of the poll frames sent form the hub to nodes is received successfully. This is why the increase of scheduled access slot length gives a clear improvement of successful received packets rate at the MAC layer.
We observe that as we increase the random access length, we have more successful received packets rate (RAP slot =2). However we cannot increase more the length of random access, because if we see in the output of the radio module that we have more interfered packets as random access length increases. The reason is that we have more opportunities for packets to collide (since CSMA is imperfect).

VII. CONCLUSION ANDFUTURE WORKS

WBANs provide promising applications in medical monitoring systems to measure specified physiological data. Our Work will contribute to improve the quality of medical assistance delivery especially in needy remote healthcare centers, where the lack of medical staff and medical equipment is a big challenge.

In this paper, we studied MAC access methods used in the IEEE 802.15.6 standard. We evaluated the performances of the improvised and scheduled access in terms of received packets, latency and packet breakdown at the MAC layer. The results demonstrated the high efficiency is obtained by using scheduled access combined to polling mechanism.

As future works, we intend to implement several enhancements at the MAC layer and to study the coexistence of WBANs and interferences issue, and how we can mitigate their impacts. In addition we intend to validate the proposal in a real world setup to assess the benefits of the solution in large scale scenarios.

REFERENCES

[1] http://www.un.org/millenniumgoals/
[2] Sensatex http://www.sensatex.com
[3] K. M. Sungmee Park and S. Jayaraman. "The wearable motherboard: a framework for personalized mobile information processing (pmip). Proceedings of 39th ACM/IEEE Design Automation Conference, pages 170–174., 2002.
[4] J. G. R. DeVaul, M. Sung and A. Pentland. "Mithril 2003: applications and architecture", 7th IEEE International Symposium on Wearable Computers, pages 4–11, 2003.
[5] J. E. T. Martin, M. Jones and R. Shenoy. "Towards a design framework for wearable electronic textiles", 7th IEEE International Symposium on Wearable Computers, pages 190– 199, 2003.
[6] Lifeguard Monitoring system, http://lifeguard.stanford.edu.
[7] S. Jiang, Y. Cao, S. Iyengar, P. Kuryloski, R. Jafari, Y. Xue, R. Bajcsy, and S. Wicker, "Carenet: an integrated wireless sensor networking environment for remote healthcare (bodynets)," in *Proc. 3rd ICST Int.Conf. on Body Area Networks*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.
[8] K. Venkatasubramanian, G. Deng, T. Mukherjee, J. Quintero, V. Annamalai, and S. S. Gupta, "Ayushman A wireless sensor network based health monitoring infrastructure and testbed.," *Springer Lecture Notes in Computer Science*, vol. 3560, pp. 406–407, 2005.
[9] Ko, J.; Lim, J.H.; Chen, Y.; Musvaloui, R.; Terzis, A.; Masson, G.; Gao, T.; Destler, W.; Selavo, L.; Dutton, R. MEDiSN: Medical emergency detection in sensor networks. ACM Trans. Embed. Comput. Syst. 2010, 10, 1–29.
[10] D. Konstantas, "The Mobihealth Project. IST Project"IST-2001-36006, European Commission: Deliverable 2.6, http://www.mobihealth.org, 2004.
[11] L. Schwiebert, S. Gupta & J. Weinmann, "Research challenges in Wireless networks of Biomedical Sensors". ACM SIGMOBILE 7/01 Rome Italy; ACM ISBN 1-58113-422-3/01/07, 2001.
[12] Timmons, N. F. AND Scalon, W. G. 2009. An adaptive energy efficient MAC protocol for the medical body area network. In Proceedings of the 1st International Conference-Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology. 587–59.
[13] http://www.citi-lab.fr/project/projets-anr/banet
[14] S. Rashwand, J. Misic, "Performance Evaluation of IEEE 802.15.6 under Non-Saturation Condition," IEEE Global Telecommunications Conference (GLOBECOM), 2011.
[15] S. Rashwand, J. Misic, V. Misic, "MAC performance modeling of IEEE 802.15.6-based WBANs over Rician-faded channels," IEEE International Conference on Communications (ICC), 2012.
[16] http://standards.ieee.org/findstds/standard/802.15.6-2012.html
[17] A. K. Jacob and L. Jacob, "Energy Efficient MAC for QoS Traffic in Wireless Body Area Network", International Journal of Distributed Sensor Networks Volume 2015 (2015), Article ID 404182 12 pages
[18] http://castalia.npc.nicta.com.au
[19] Zuniga, M. and Krishnamachari, B. "Analyzing the transitional region in low power wireless links" Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004.
[20] A. Bouayad, N. Chaoui, M. El Ghazi , "Modeling and Simulation of a Wireless Body Area Network for Monitoring Sick Patient Remotely " (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 580-585.
[21] SMA-WiBAN, "MAC and Security Baseline Proposal", IEEE 802.15 Documents, Document no. 196, rev.2, Mar 17th 2010 2009.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Comparative Analysis based Classification of KDD'99 Intrusion Dataset

Shashikant Upadhyay
PG Scholar, Department of Computer Science & Engg, VNSIT, Bhopal, INDIA

Rajni Ranjan Singh
Assistant Professor, Department of Computer Science & Engg, VNSIT, Bhopal, INDIA

***Abstract*— Classification play as important role for constructing Intrusion Detection System (IDS) which is used for classifies network traffic information. Mostly traffic information is classified on the basis of normal or anomaly behaviors. In this work an attack specific classification of KDD cup dataset is proposed. Here Naive Bayes, j48, j48 craft, Bayes net and random forest classifiers are explored for the classification. A comparative analysis is carried out to identify best classifiers for each individual attacks based on Precision, recall, f-measures and ROC Curve Area performance criteria's. It is observed that random forest is the best classifier among all used methods. However despite of the high classification of random forest, BayesNet performed better in context of false positive rate of classification. Similarly ROC Curve Area rate of Bayes net is the best among the all used methods.**

Keywords- Classification; KDD Cup '99; Naive Bayes; BayesNet J48; J48graft and Random Forest.

I. INTRODUCTION

With the huge growth of the usage of computer and the enormous rise in the number of applications and services running on top of it, network security is becoming increasingly more complex and vital. Nagunwa et. al. [1] year of 2012, 556 million consumers were victimized by more than 30 million cyber attack activities. Hence it is paramount important to research and adopt method for the intrusion detection. Therefore the task of Intrusion Detection Systems (IDS's) is applied identify attacks and anomalies within the network. The KDD Cup'99 dataset has been used by a large amount of the researchers as a test bed for the improvement of IDS and IPS system.

The organization of the rest of the paper is as follows: Section 2 describes the related work in this area. In Section 3 include detailed description of KDD Cup "99. Section 4 presents about proposed work. Section 5 explains the Observation and findings.

II. RELATED WORK

Asak et al. [2] proposed a method for discriminant analysis of Machine learning based Intrusion Detection. In which a feature selection based method is utilized for the classification of individual attack. Author's utilizes system log information as experimental purpose.

Ramani et. al. [3] proposed a Discriminant Analysis based Feature Selection of KDD Intrusion Dataset. In this paper [3], important features of KDD Cup 99 attack dataset are extracted by the use of discriminant analysis method. Author's mentioned that proposed method is suffering by two- class classification or multiclass classification problems.

Kayacik et. al. [4] proposed a work of feature relevance analysis on KDD'99 dataset on the basis of information gain. Feature relevance is expressed in terms of information gain, which gets higher as the feature gets more discriminative. On the basis of result authors sagest that normal, neptune and smurf classes are highly related to certain features that make their classification easier. On the other hand authors told about certain features have no contribution to intrusion detection.

Balakrishnan et. al[5] proposed a new feature selection algorithm based on Information Gain Ratio. The feature selection decreases the classification time. The author claims that proposed IDS reduce the false positive rates and classification time.

III. KDD CUP '99 DATASET DESCRIPTION

The KDD Cup '99 dataset developed by MIT Lincoln Laboratory in the year 1999 known as KDD CUP '99 Intrusion Detection Dataset. The KDD CUP '99 training dataset is a collection of 494,020 records. Every dataset tuple is a single connected vector described through 41 characteristic values and exactly one label of either 'normal' or an 'attack'. The size of KDD'99 is 50MB. In KDD'99 dataset every attack is classified into exactly one of the following four categories:

I. **Denial of Service Attack (DO**S): In this class of attack the attacker makes an effort to create a resource unavailable for its authorized users.DOS attacks are very common in massive booking and banking systems where unavailability of a resource will cause serious economic losses.

II. **User to Root Attack (U2R):** This is one of the most dreaded cyber-attacks in the industry. Here the attacker gains access to an end user's account and tries to obtain root access to the system. After gaining access, the attacker can then disrupt the normal functioning of the entire system.

III. **Remote to Local Attack (R2L):** The main purpose of this type of attack is to get local access to an end user machine.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The attacker sends a packet to a machine over a network and then tries to get access to an account on that machine.

IV. **Probing Attack (PROBE):** In this category of attacks attacker performs snooping of targets networks in order to get live system, open ports etc. this information is used by the attacker for vulnerabilities assessment.

A. Classifiers Used

1. **Naive Bayes classifier: -** A naive Bayes classifier is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable [6].

Naïve Bayes classifier assumes that the effect of the value of a predictor (X) on a given class (C) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P\left(\frac{C}{X}\right) = \frac{P\left(\frac{X}{C}\right)P(C)}{P(X)}$$

P(C/X) is the posterior probability of class (target) given predictor (attribute). P(C) is the prior probability of class. P(X/C) is the likelihood which is the probability of predictor given class. P(X) is the prior probability of predictor.

2. **BayesNet: -** A Bayesian network is a graphical model for probabilistic relationships among a set of variables. Nodes represent random variables they may be observable quantities, latent variables, unknown parameters or hypotheses. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node [7].

3. **J48 Classifier: -** A J48 Classifier is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

4. **J48 Graft: -** J48graft produces a grafted Decision Tree from a J48 tree. The graft method adds nodes to an existing decision tree with the aim of reducing prediction errors. These algorithms identifies regions of the instance space that don't seem to be occupied by training instances, or occupied solely by misclassified training instances, and take into account various classifications for those regions.

5. **Random forests Classifier:-**Random forests are an ensemble learning technique for classification, regression and extra tasks, that operate by constructing a large number of decision trees at training time and outputting the category that's the mode of the categories (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of over fitting to their training set.

B. Performance Parameter for Evaluation

1. **True positive (TP)/ Recall :** It is the proportion of positive cases that were correctly classified as positive, as calculated using the equation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

2. **False positive (FP):** It is the proportion of negative cases that were incorrectly classified as positive, as positive, as calculated using the equation:

$$\text{Recall} = \frac{FP}{TN + FP}$$

3. **Precision: -** Accuracy (i.e., Precision) is the proportion of the total number of attacks that are correctly detected. It is determined using the equation:

$$\text{Accuracy} = \text{Precision} = \frac{TP}{TP + FP}$$

Here, TP is True Positive; FP is False Positive, FN is False Negative.

4. **F- Measure: -** F- measure that mixes precision and recall is the harmonic mean of precision and recall is known as F-measure.

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. **ROC: -** Receiver Operator Characteristics (ROC) illustrates the tradeoff between sensitivity and specificity. ROC curves plot the true positive rate vs. the false positive rate, at varying threshold cutoffs.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. PROPOSED WORK

Proposed work is consist of two main modules, a dataset preparation and second is classification module; description of each module is given here.

A. Dataset preparation:

Training and testing dataset is divided into the individual attack nominals. By defaults KDD dataset is composed of 5 attack categories that are Normal, DOS, R2L, U2R, and Probe however in this work dataset is processed and instead of over mention 5 attack categories dataset is divided into 22 attack sub categories shown in the Table I.

Table I. Number of Individual Attack in the Kddcup'99 test set and distribution of attacks.

Attack Name	Attacks in training dataset	Category
normal.	97,277	NORMAL
neptune.	101,201	DOS
back.	2,203	DOS
teardrop.	979	DOS
smurf.	280,790	DOS
pod.	264	DOS
land.	21	DOS
warezclient.	1,020	R2L
multihop.	7	R2L
ftp_write.	8	R2L
imap.	12	R2L
guess_passwd.	53	R2L
warezmaster.	20	R2L
spy.	2	R2L
phf.	4	R2L
buffer_overflow	30	U2R
loadmodule.	9	U2R
rootkit.	10	U2R
perl.	3	U2R
portsweep.	1,040	PROBE
satan.	1,589	PROBE
ipsweep.	1,247	PROBE
nmap.	231	PROBE

B. Classification

Processed dataset is applied to the Naive Bayes, Bayes net, J48, J48graft, and Random forest classifiers. An experiment setup is built consist of Intel Core i3 Processors, 4 GB RAM, 500 GB HDD, Ubuntu 14.10 Operating System and Weka machine learning workbench is utilized for the classification task.

V. OBSERVATION AND FINDINGS

Classification is performed based on the 22 attack categories of testing dataset. In context of True Positive rate (recall) the random forest methods detect 14 attack categories with highest TP Rate. This is the best among all used methods. However buffer_overflow, pod and snmpgetattack attacks are more accurately detected by BayesNet classifier. Similarly best true positive rate of warezmaster attack is given by the J48 graft method. ps attack is correctly detected by Naive Bayes which is shown in Table II.

Table II. Comparative Study of Five Classifiers for Individual Attacks on "TP Rate/ Recall" Score.

S.No.	Attacks Name	Naive Bayes	BayesNet	J48	J48Graft	Random Forest
1	apache2	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**
2	back	**1**	**1**	0.995	0.995	**1**
3	buffer_overflow	0.714	**1**	0.714	0.571	0.714
4	guess_passwd	0.844	0.993	0.994	0.994	**1**
5	httptunnel	0.949	**0.983**	0.881	0.881	**0.983**
6	ipsweep	0.968	**0.979**	**0.979**	**0.979**	**0.979**
7	land	0	**1**	**1**	**1**	**1**
8	mailbomb	0.976	**1**	**1**	**1**	**1**
9	mscan	0.978	0.995	0.986	0.986	**1**
10	neptune	0.995	0.996	**1**	**1**	**1**
11	nmap	**1**	**1**	**1**	**1**	**1**
12	normal	0.636	0.797	0.947	0.947	**0.948**
13	pod	0.914	**0.971**	0.943	0.943	0.943
14	portsweep	0.904	0.968	0.92	0.928	**0.976**
15	processtable	0.98	0.996	0.996	0.992	**1**
16	ps	**0.778**	0.444	0.111	0.111	0.333
17	saint	0.02	0.746	**0.93**	**0.93**	0.922
18	satan	0.973	0.927	**0.976**	0.974	0.958
19	smurf	0.999	**1**	**1**	**1**	**1**
20	snmpgetattack	0.606	**0.992**	0.654	0.654	0.65
21	snmpguess	0.971	**0.996**	**0.996**	**0.996**	**0.996**
22	warezmaster	0.442	0.998	0.988	**0.989**	0.993
Weighted Avg.		0.909	0.958	**0.98**	**0.98**	**0.98**

In context of false positive rate Naive Bayes generate highest false alert and BayesNet generate least false alerts. However Naive Bayes and BayesNet do not falsely detect normal traffic as suspicious, shown in Table III.

In context of Precision score, Random forest is the best classifiers it gives highest precision rate for 16 different attacks out of 22 arrack categories. However highest precision score of normal and Apache2 attacks is given by Naïve Bayes and BayesNet respectively. Similarly best precision rate of buffer_overflow attack is given by the j48 graft method which is shown in Table IV.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table III. Comparative Study of Five Classifiers for Individual Attacks on "FP Rate" Score.

S.No.	Attacks Name	Naive Bayes	BayesNet	J48	J48Graft	Random Forest
1	apache2	0	0	0	0	0
2	back	0	0	0	0	0
3	buffer_overflow	**0.003**	0.001	0	0	0
4	guess_passwd	**0.002**	0	0	0	0
5	httptunnel	**0.001**	0	0	0	0
6	ipsweep	**0.004**	0.001	0	0	0
7	land	0	**0.001**	0	0	0
8	mailbomb	**0.001**	0	0	0	0
9	mscan	**0.001**	0	0	0	0
10	neptune	0	0	0	0	0
11	nmap	0	0	0	0	0
12	normal	**0**	0	**0.011**	**0.011**	**0.011**
13	pod	**0.004**	0	0	0	0
14	portsweep	0	0	0	0	0
15	processtable	0	0	0	0	0
16	ps	**0.013**	0	0	0	0
17	saint	**0.002**	0	0	0	0
18	satan	**0.002**	0.001	0	0	0
19	smurf	**0.005**	0	0	0	0
20	snmpgetattack	0.025	**0.036**	0.01	0.01	0.01
21	snmpguess	**0.031**	0.001	0	0	0
22	warezmaster	**0.001**	**0.001**	0	0	0
Weighted Avg.		**0.004**	0.001	0.002	0.002	0.002

Table IV. Comparative Study of Five Classifiers for Individual Attacks on "Precision" Score

S.No.	Attacks Name	Naive Bayes	BayesNet	J48	J48Graft	Random Forest
1	apache2	0.861	**1**	0.982	0.982	0.996
2	back	0.992	0.995	0.997	**1**	**1**
3	buffer_overflow	0.014	0.097	0.417	**1**	0.5
4	guess_passwd	0.876	0.999	0.996	0.998	0.999
5	httptunnel	0.487	0.644	0.963	0.981	**1**
6	ipsweep	0.196	0.514	**1**	**1**	0.989
7	land	0	0.031	**1**	**1**	**1**
8	mailbomb	0.943	**1**	0.998	0.996	**1**
9	mscan	0.756	0.948	0.989	0.984	**1**
10	neptune	**1**	**1**	**1**	**1**	**1**
11	nmap	**1**	0.813	**1**	**1**	**1**
12	normal	**0.999**	0.998	0.953	0.953	0.954
13	pod	0.071	0.567	0.868	0.892	**0.971**
14	portsweep	0.79	0.818	0.991	0.991	**1**
15	processtable	0.992	**1**	**1**	**1**	**1**
16	ps	0.005	0.211	0.333	**1**	**1**
17	saint	0.026	0.788	**0.95**	**0.95**	0.93
18	satan	0.71	0.884	0.964	0.966	**0.978**
19	smurf	0.996	**1**	**1**	**1**	**1**
20	snmpgetattack	0.383	0.411	**0.618**	**0.618**	**0.618**
21	snmpguess	0.195	0.928	0.999	0.999	**1**
22	warezmaster	0.748	0.804	0.989	0.988	**0.991**
Weighted Avg.		0.965	**0.981**	**0.981**	**0.981**	**0.981**

Observation of F-Measure rate is very much similar to the over mentioned Precision observations which is shown in Table V.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table V. Comparative Study of Five Classifiers for Individual Attacks on "F- Measure" Score.

S.No.	Attacks Name	Naive Bayes	BayesNet	J48	J48Graft	Random Forest
1	apache2	0.924	**0.998**	0.989	0.993	0.996
2	back	0.996	0.997	0.996	0.997	**1**
3	buffer_overflow	0.028	0.177	0.526	**0.727**	0.588
4	guess_passwd	0.859	0.996	0.995	0.996	**0.999**
5	httptunnel	0.644	0.779	0.92	0.929	**0.991**
6	ipsweep	0.326	0.674	**0.989**	**0.989**	0.984
7	land	0	0.06	**1**	**1**	**1**
8	mailbomb	0.959	**1**	0.999	0.998	**1**
9	mscan	0.853	0.971	0.988	0.985	**1**
10	neptune	0.998	0.998	**1**	**1**	**1**
11	nmap	**1**	0.897	**1**	**1**	**1**
12	normal	0.777	0.886	0.95	0.95	**0.951**
13	pod	0.132	0.716	0.904	0.917	**0.957**
14	portsweep	0.843	0.886	0.954	0.959	**0.988**
15	processtable	0.986	0.998	0.998	0.996	**1**
16	ps	0.01	0.286	0.167	0.2	**0.5**
17	saint	0.023	0.766	**0.94**	**0.94**	0.926
18	satan	0.821	0.905	**0.97**	**0.97**	0.968
19	smurf	0.997	**1**	**1**	**1**	**1**
20	snmpgetattack	0.469	0.582	**0.636**	**0.636**	0.634
21	snmpguess	0.324	0.961	**0.998**	**0.998**	**0.998**
22	warezmaster	0.556	0.89	**0.989**	**0.989**	0.992
Weighted Avg.		0.926	0.964	0.98	0.98	**0.981**

ROC Curve Area rate of Naïve Bayes is the lowest and remaining four classifier having same average rate. It is interesting to see that highest ROC Curve rate of buffer_overflow and httptunnel is given by BayesNet classifier which is shown in Table VI.

Table VI. Comparative Study of Five Classifiers for Individual Attacks on "ROC Area" Score.

S.No.	Attacks Name	Naive Bayes	BayesNet	J48	J48Graft	Random Forest
1	apache2	0.999	**1**	0.998	0.998	**1**
2	back	**1**	**1**	**1**	0.997	**1**
3	buffer_overflow	0.975	**1**	0.928	0.856	0.929
4	guess_passwd	0.994	**1**	0.997	0.997	**1**
5	httptunnel	0.997	**0.998**	0.961	0.942	0.992
6	ipsweep	0.988	0.993	0.994	0.994	**0.998**
7	land	**1**	**1**	**1**	**1**	**1**
8	mailbomb	0.999	**1**	**1**	**1**	**1**
9	mscan	**1**	**1**	0.997	0.999	**1**
10	neptune	0.999	**1**	**1**	**1**	**1**
11	nmap	**1**	**1**	**1**	**1**	**1**
12	normal	0.98	0.997	**0.998**	**0.998**	**0.998**
13	pod	0.976	**1**	0.971	0.971	**1**
14	portsweep	0.999	**1**	0.988	0.988	**1**
15	processtable	0.998	**1**	0.998	0.998	**1**
16	ps	0.992	**0.997**	0.886	0.83	0.944
17	saint	0.969	**0.999**	0.992	0.988	0.988
18	satan	0.999	**1**	0.996	0.994	0.994
19	smurf	0.997	**1**	**1**	**1**	**1**
20	snmpgetattack	0.977	0.989	**0.99**	**0.99**	**0.99**
21	snmpguess	0.993	**1**	0.999	0.999	**1**
22	warezmaster	0.997	**1**	0.997	0.998	**1**
Weighted Avg.		0.994	**0.999**	**0.999**	**0.999**	**0.999**

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

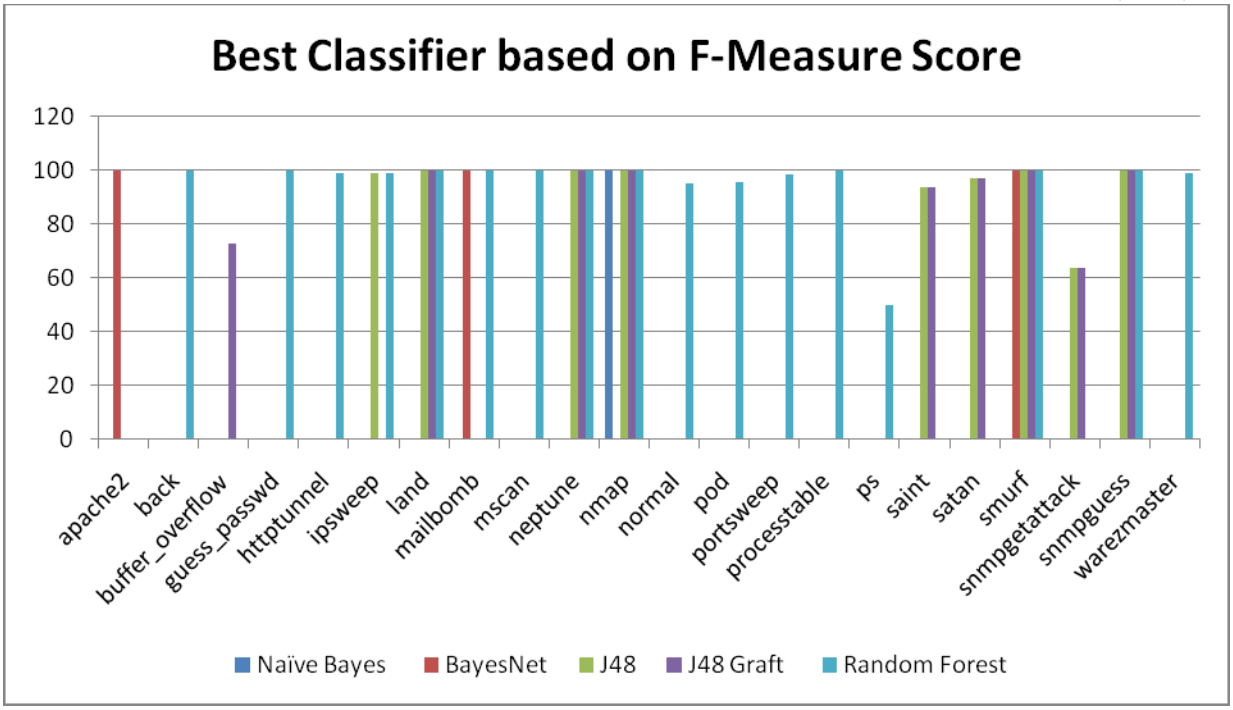


Figure1. Best Classifier based on F-Measure Score

Table VII. Best Classifier based on F-Measure Score

Sl. No.	Attack Name	Percentage in %	Best F-Measure Classifier
1	apache2	99.8	BayesNet
2	back	100	Random Forest
3	buffer_overflow	72.7	J48 Graft
4	guess_passwd	99.9	Random Forest
5	httptunnel	99.1	Random Forest
6	ipsweep	98.9	J48, Random Forest
7	land	100	J48, J48 Graft ,Random Forest
8	mailbomb	100	BayesNet, Random Forest
9	mscan	100	Random Forest
10	neptune	100	J48, J48 Graft ,Random Forest
11	nmap	100	Naïve Bayes , J48, J48 Graft ,Random Forest
12	normal	95.1	Random Forest
13	pod	95.7	Random Forest
14	portsweep	98.8	Random Forest
15	processtable	100	Random Forest
16	ps	50	Random Forest
17	saint	94	J48, J48 Graft
18	satan	97	J48, J48 Graft
19	smurf	100	BayesNet, J48, J48 Graft ,Random Forest
20	snmpgetattack	63.6	J48, J48 Graft
21	snmpguess	99.8	J48, J48 Graft ,Random Forest
22	warezmaster	99.2	Random Forest

VI. CONCLUSION

This work proposes an attack specific classification of KDD cup dataset. Here instead of five broad categories, classification is carried out for 22 attack subcategories. A comparative analysis is performed in order to identify best classifiers among 5 classifiers (Naive Bayes, Bayes net, J48, J48graft, and Random forest).

Following conclusions are drawn from the work.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

a) Instead of using single classifier for all attack categories, it is beneficial to utilize individual classifier for each category of attack, for example, In order to detect apache2 attack, BayesNet is the best classifiers because it gives best result for all measurement criteria's.

b) Overall Random forest is the best classifiers it gives highest F-Measures for 16 different attack categories out of 22, shown in figure 1 and Table VII.

c) Naive Bayes generate highest false positive rate.

d) Naive Bayes gives the lowest ROC Curve area rate.

REFERENCES

[1] Thomas Nagunwa," Behind Identity Theft and Fraud in Cyberspace: The Current Landscape of Phishing Vectors" International Journal of Cyber-Security and Digital Forensics (IJCSDF, ISSN: 2305-0012)-2014, 3(1): pages 72-83.

[2] Midori Asak a, Takefumi Onabura, T adashi Inoue, Shigeki Goto. 2002. Remote Attack Detection Method in IDA: MLSI-Based Intrusion Detection using Discriminant Analysis. Proceedings of the 2002 Symposium on Applications and the Internet (SAINT.02), IEEE.

[3] Dr. R.Geetha Ramani, Dr.S.Siva Sathya, K.Sivaselvi. 2011. Discriminant Analysis based Feature Selection in KDD Intrusion Dataset, International Journal of Computer Applications (0975 – 8887) Volume 31– No.11, October 2011.

[4] H. Günes Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood, Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets.

[5] Senthilnayaki Balakrishnan, Venkatalakshmi K, Kannan A. 2014. Intrusion Detection System Using Feature Selection and Classification Technique. International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 4, November 2014.

[6] K. Huang (2003) Discriminative Naive Bayesian Classifiers, Department of Computer Science and Engineering, The Chinese University of Hong Kong, 1-21

[7] http://www.fit.vutbr.cz/study/courses/VPD/public/0809VPD-Vanek.pdf

AUTHORS PROFILE

Shashikant Upadhyay

Shashikant Upadhyay received his B.E. degree in Information Technology from Bansal College of Engineering Mandideep under Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal in 2010 and pursuing M Tech. (Regular) degree in Computer Science & Engineering from VNS Group of Institutions, Faculty of Engineering (Formerly known as VNS Institute of Technology) , Bhopal., batch 2012-2014. His research interests include improvement of Intrusion Detection System in Data Mining.

Mr. Rajni Ranjan Singh

Rajni Ranjan Singh, working as Assistant Professor in Department of Computer Science & Engineering at VNS Group of Institutions, Faculty of Engineering (Formerly known as VNS Institute of Technology) , Bhopal. His research activities are based on digital forensics, data mining and network security.

Policy-based smart adaptive quality of service for network convergence

Ayoub BAHNASSE
Dept. Physic, Lab STIC, Faculty of Sciences EL Jadida
University Chouaïb Doukkali
El Jadida, Morocco

Najib ELKAMOUN
Dept. Physic, Lab STIC, Faculty of Sciences EL Jadida
University Chouaïb Doukkali
El Jadida, Morocco

***Abstract*—the management of a convergent network has become one of the primordial and challenging tasks. Taking into account the rapid evolution of networks' size and the diversity of deployed applications, and their requirements in terms of quality of service (QoS), it is quite difficult, on the one hand, to ensure through a manual customization of Quality of Service policies a level of performance required by some flows, on the other hand, to automatically and dynamically adapt QoS policies to instantly optimize the resources, in order to guaranty for applications the required bandwidth and to ensure low loss rates , latency and jitter.**

This paper presents a novel model of the policy-based Smart adaptive Quality of Service management for clients and Internet service providers. The model proposes separation and distribution of: Data, control and management plans, for a quick and effective treatment, as well as automatically and dynamically reallocates bandwidths, reduce the loss rate and minimize the delay and jitter in a converged network.

Keywords- Quality of service, Adaptive QoS, Smart QoS, Policy-based, network convergence.

I. INTRODUCTION:

The Internet Protocol (IP) route all the packets as quick as possible within its capabilities (Best Effort), with no guarantee of neither bandwidth, neither jitter delay nor loss rate. The (Best Effort) principle works well as long as applications are less demanding in terms of resources, however, with the appearance of Network Convergence [1], establishing the quality of service mechanisms have become essential to ensure and reserve for applications the required resources for their smooth functioning. Quality of service functionalities have as a primary objective to overcome the Internet Protocol limits under two models, Differentiated Service model and Integrated Service model.

Differentiated Service "DiffServ" [2] is a mechanism that allows classifying data in many Behavior Aggregate "BA" based on the DS (DiffServ) field of the IPv4 packet, Classification, control and flow labeling are performed by the "Edge Router", core routers treats packets based on the value of the DS field according to a specific behavior known as Per Hop Behavior "PHB"; two PHB behaviors are defined;

Expedited Forwarding "EF" [3] which aims to guarantee bandwidth with a low loss rate, low delay and low jitter; Assured Forwarding "AF" [4] this family is split into four different classes ensuring a minimum delay and bandwidth, each class contains three priority levels (Drop Precedence).

Integrated Service "IntServ", Contrary to the DiffServ model, it doesn't depend on the Internet Protocol functioning, it's based on the RSVP (Resource ReSerVation Protocol) signaling, RSVP allows routers to reserve a bandwidth for a particular flow across a giving data path, this signaling protocol doesn't have just as a task to accomplish the Quality of Service, but can used in the MPLS Networks (Multi-Protocol Label Switching) [5]. The IntServ model defines two services, Guaranteed Service (GS) [6] and Controlled Load (CL) [7], the GS service gives a strict guarantee of the maximum delay of each packet and ensures zero loss rates, this service is invoked by the transmitter that specifies its requirements, In other words, its traffic Settings (Tspec) and then by the receiver by indicating a desired service level (Rspec), the CL service proposes an end to end packet treatment closer than Best Effort service, except of the latency and the packets loss rate won't exceed a lot the minimum values achieved by packets in the same lightly loaded network, the transmitter sends a Tspec message, specifying an estimation of traffic to generate, not traffic settings.

The management of quality of service on modern networks has become an important but complex task, given that a variety of flow can pass through the same network, each flow has its own QoS requirement depending on its nature, several research works has been done for the Quality of Service management, by an autonomous or even server-based approaches, the [8-11] works relate to the topic: Policy Based Network Management (PBNM) of the Quality of Service, but proposed models are addressed to architectures with a limited class or virtual machine number, as well as the validation tests were not performed in converged networks in unfavorable jitter, latency and high loss rate requirements, the works [12-15] define autonomous management of QoS approach, based on a data plan, a single control plan and a management plan, the concept of logical layers, provided more flexibility, scalability and modularity in terms of automatic adjustment of QoS settings, [16,17] proposes distributed approaches of several control plan, these approaches overcome the deficiencies of [12-15] works that uses a single control plan for simultaneous management of multiple functionalities (Routing, QoS, traffic engineering…) which can cause additional latency on applications to manage, [18] it proposes an adaptive management Framework of network service architectures, but without using the concept of control and data plans.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The discussions on related works were an incentive to design and implement a new distributed, scalable, flexible, adaptive model called: Policy-Based Smart Adaptive Quality Of Service For network convergence (PB SAQOS) to:

- Use a server with distributed data, control and management plans, to ensure efficient, fast and automatic adjustment management of QoS policies, on both, client and service provider edges.
- Use a new adaptive algorithm of additional resources reservation or release.

The rest of the paper will be organized as following; the 2nd section presents the PB SAQOS and its different modules, section 3 discuss adaptation and resources release algorithms, the 4th section will be reserved for the simulation and model evaluation, and we will conclude in 5th Section.

II. THE PB SAQOS MODEL:

As illustrated in Figure 1, the architecture of the PB SAQOS model is composed by three logical layers fully distributed: Data Layer, Control Layer and Management Layer. Each layer contains three plans, respectively: Data Plan (Section II-A), Control Plan (Section II-B) and Management Plan (Section II-C).

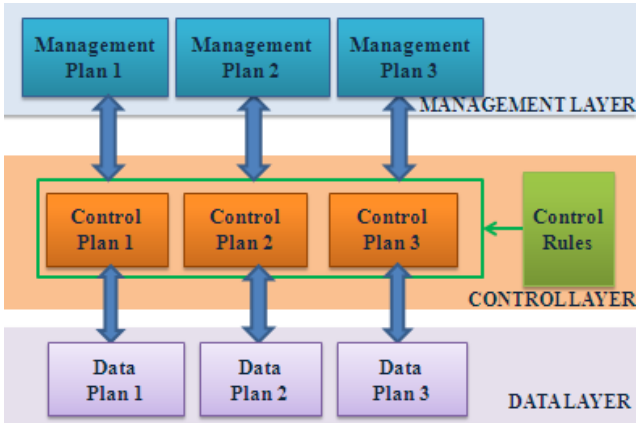


Fig. 1 Overview of PB-ASQOS model

Data Layer contains multiple data plans, which are responsible for the automatic detection of the topology and a QoS policy delivery using a secure Remote Method Invocation (RMI) connection. [19]

Control layer includes several control plans that ensure admission control, adaptive control, definitions of thresholds triggering actions, classification criteria and statistics collection. For reasons of scalability and modularity, the definition of thresholds and criteria classifications are centralized in the "Control Rules" module, these will be applied automatically and dynamically to the control plan.

Management Layer, containing several management plans, responsible of the execution of adaptation (Section III-A), and resources release (Section III-B), in order to dynamically adjust QoS policies.

Figure 2 shows the internal architecture of each plan,

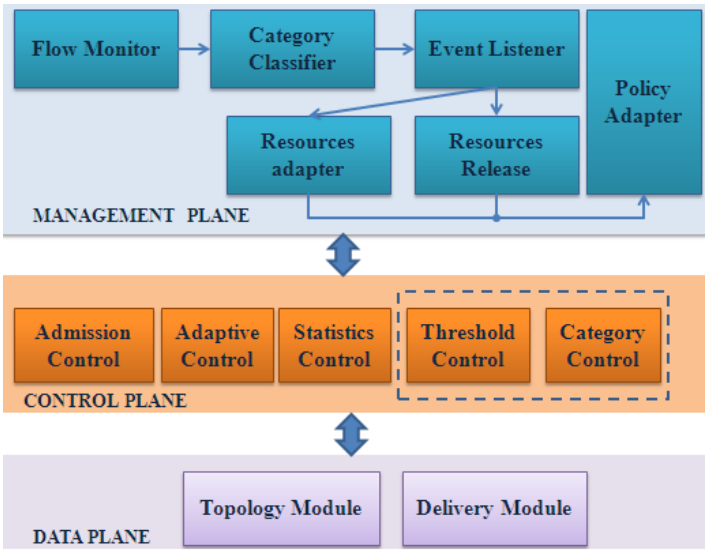


Fig. 2 PB SAQOS Architecture plans

A. *Data plan:*

The data plan is composed of two modules: Topology Module and Delivery Module:

Topology Module: this module performs an automatic and a dynamic detection of routers of the topology, in the case of Dynamic Multipoint Virtual Private Network (DMVPN) architectures, the data plan ensures: the dynamic consultation of the NHRP cache of the HUB router in order to identify equipment of the architecture, for other networks, the topology detection can be performed by a manual specification of an single or range of IP addresses of equipment to control.

Delivery Module: This module uses RMI to open secure channels to the detected routers and send new policies modification, the choice of the RMI has been done due to its flexibility, task parallelism and the real-time processing [20].

B. *Control plan:*

The control plan is composed of five modules: Threshold Control, Category Control, Statistics Collector, Admission Control, and adaptive control, the first two modules can be set by the administrator (optional).

Threshold Control: this module allows defining thresholds that trigger adaptation events or resources' release. Thresholds to define are: the used bandwidth, the loss rate, the latency and the tolerated jitter:

- Adaptation metric: loss rate, latency and jitter default-values by the model are respectively: 1%, 15ms, and 150ms.
- Release metric: loss rate, latency and jitter default-values by the model are respectively: 0.5%, 10ms, and 60ms.

Category Control: the module allows defining three categories and their belonging criteria, the categories contain several classes with non-specified numbers, the Critical Excess (CE) category, the No Excess (NE) category and the Extrem No Excess (ENE) category.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The CE category contains classes of which: the used bandwidth ≥ 70%, loss rate > 1%, jitter ≥15ms and latency ≥ 150 ms,
- The NE category contains classes of which: the used bandwidth is between 40% and 70%,
- The ENE category contain classes of which: the used bandwidth is less than 40%.

Statistics Control: This module defines the set of metrics to monitor using an active metrology, such as:

- The amount of the allocated bandwidth,
- The amount of consumed bandwidth,
- The amount of the unused bandwidth,
- the loss rates per class, per device, per flows and per interface,
- the latency per class, per device, per-flow and per interface for real-time flows,
- Jitter per class, per device, per-flow and per interface for real-time stream.

Adaptive Control: This module receives additional reservation requests or resources' release of classes from the Management Plan, and decides either to allow or to reject additional reservation or release requests of resources depending on: their belonging categories, the total free bandwidth to allocate, the total number of classes in CE category and class priority soliciting more resources.

Admission Control: This module operates at service provider side, it takes into account client's resource reservation requests and available capacity to determine whether to accept or not a QoS request, this module include additional aspects such as; monitoring and controlling network resources based on policies derived from clients such as: the identity of the requester, application bandwidth requirement.

C. Management plan:

The management plan defines the set of functions performed after validation of the Adaptive Control module; this plan is responsible of the intelligent adaptation or release of resources in terms of defined categories in the control plan. The management plan is made up of 6 modules: Flow monitor, Category classifier, Event Listener, Resources Adapter, Resources Release and Policy Adapter.

Flow Monitor: this module allows:

- To automatically determining which among the metrics in the Statistics Control module will be used, depending on the class flow nature.
- To detect periodically flows statistics depending on the metrics previously chosen.

Category Classifier: this module allows affecting a class to a defined category in the Category Control module of the Control plan, the dividing of classes to categories provides more flexibility and scalability, the treatment will be precisely performed on specific classes of the CE category on the one hand, and on the other, a category can contain several classes, that doesn't require a change of the model's architecture.

Event Listener: this module detects categories changes of classes, in order to decide which algorithm to run:

- For a negative change, which means a switch to the CE category, the Event Listener runs the adaptation algorithm,
- For a positive change, this means a switch to the NE or ENE categories, the Event Listener runs the resources' release algorithm,

Resource Adapter: this module allows running an adaptation algorithm of resources detailed in the in the Section III-I.

Resource Release: this module allows running a release algorithm of resources detailed in the in the Section III-II.

Policy Adapter: depending on decisions made by all the previous modules, this module prepares new policies to deliver by the Delivery module of the Data Plan.

III. DECISIONAL ADAPTATION AND RELEASE OF RESOURCES ALGORITHM:

This section describes the two modules:

- The adaptation of resources. used by the resource adapter module of the management plan,
- Decisional release of resources. Used by the Resource Release module of the management plan.

A. Adaptation of resources :

This component provides multiple additional bandwidth allocations to classes belonging to the EC category.

- Function 1 – collects unused bandwidths:

Classes of the NE and ENE categories will allocate 50% of their unused bandwidths (Eq. 1)

$$BWoffered\ Classi\ _{i=1}^{n} = BWa\ Classi - BWu\ Classi \times 0.5 \quad (1)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

n : numbers of classes in the CE category.

BWoffered : Offered bandwidth by a class « i » of a specific category.

BWa : allocated bandwidth by a class.

BWu : used bandwidth by a class.

$$BWoffered\ Total\ NE = \sum_{i=1}^{j} BWoffered\ Classi \quad 2$$

$$BWoffered\ Total(ENE) = \sum_{i=1}^{j} BWoffered\ Classi \quad (3)$$

Equation (2) represents total bandwidth offered by all the classes (number=j) of the NE category.

Equation (3) total bandwidth offered by all the classes (number=j) of the ENE category.

- Function 2 – Division of used bandwidths :

After collecting the bandwidth of the NE and ENE categories (Function 1), the division of these resources can be performed according to two possible scenarios:

Scenario 1: if classes of the CE category have the same priority, then an equal cost division of collected resources has to be performed, the correspondent algorithm is as follows:

Step 1/3: Calculate the dynamic bandwidth offered by the NE category, allocated for each class of the CE category (Eq.4).

$$BWallocated1 = BWoffered\_total\ NE\ /\ Classes(CE) \quad (4)$$

Step 2/3: Calculate the dynamic bandwidth offered by the ENE category, allocated for each class of the CE category (Eq.4).

$$BWallocated2 = BWoffered\_total\ ENE\ /\ Classes(CE) \quad (5)$$

Step 3/3: Calculate the new dynamic bandwidth of the CE category classes (Eq.6), by performing a summing of the current bandwidth with determined bandwidths by the equations 4 and 5.

$$BWD\ Class\ i \sum_{i=1}^{n} = BWallocated1 + BWallocated2 + BWa\ Class\ i \quad (6)$$

Where, "n" represents classes' number of the CE category.

Scenario2: if classes of the CE category have different priorities, then an unequal cost division depending on the priority must be performed.

Step 1/4: calculate priorities' average value of the CE class (Eq.7),

$$AVGPriority = priorities\ (CE)\ /\ Classes\ CE \quad (7)$$

Step 2/4:

1. Calculate the reference bandwidth of the NE category "RefBW1"(Eq.8)

$$RefBW1 = (AVGPriority \times BWoffered\_total(NE))/100 \quad (8)$$

2. Calculate the reference bandwidth of the E NE category "RefBW2"(Eq.9)

$$RefBW2 = (AVGPriority \times BWoffered\_total(ENE))/100 \quad (9)$$

Step 3/4:

1. Calculate the allocated bandwidth "BWallocated1" by the NE category to classes of the ENE category depending on their priorities "Priority (Class)", (Eq.10).

$$BWallocated1 = (AVGPriority \times BWoffered\_total(NE))/100 \quad (10)$$

2. Calculate the bandwidth allocated by the ENE category "BWallocated2" to classes of the ENE category depending on their priorities "Priority (Class)", (Eq.11).

$$BWallocated2 = (AVGPriority \times BWoffered\_total(ENE))/100 \quad (11)$$

Step 4/4: Calculate the new dynamic bandwidth of classes of the CE category (Eq.12), by adding the current bandwidth to the determined bandwidths on the equations 10 and 11.

$$BWD\ Class\ i \sum_{i=1}^{n} = BWallocated1 + BWallocated2 + BWa\ Class\ i \quad (12)$$

The figure3 illustrates a synthesis of the resources' adaptation algorithm.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

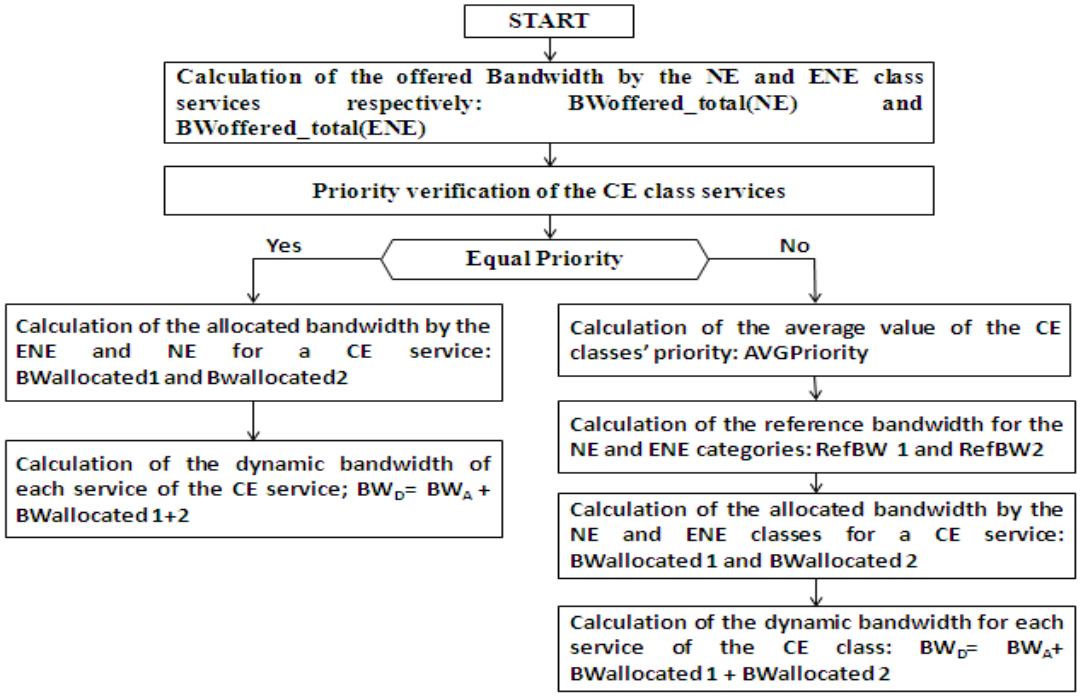


Fig. 3 Resources Adaptation diagram

B. Decisional release of resources:

This component is responsible of the decisional resources' release, the release process works on three phases:

1. Monitor the category of the donor classes of resources.
2. Collect state information of donor classes
3. Verify offered resources by the donor class in the adaptation phase.
4. Verify by interviewing the Admission Control module, the quantity of resources that the donor class can recover if needed.
5. Verify by interviewing the adaptation Control module, the beneficiary classes that have to release the allocated resources, the decision is made depending on priorities, traffic nature and the selected metrics by the Flow Monitor module.
6. Verify by interviewing the adaptation Control module, the quantity of resources that the beneficiary class can return.

For the rest of the algorithm, we will present the Beneficiary Classes as ″BC″, and the Donor Classes by ″DC″.

The algorithm runs when the DC reaches the release thresholds defined by the Threshold Control.

If DC reaches the thresholds previously defined, the following algorithm will be executed on the non-priority BC compared to DC:

Start

If the BC priority is less than that of DC, so BC will return 50% of the bandwidth that it had allocated "BWallocated". then verify the used bandwidth by DC,

If the used bandwidth is less than 60%, then end the process,

If not, BC will return 25% of the remaining "BWallocated" bandwidth,

If the used bandwidth by DC is less than 60%, then end the process,

If not BC will return 25% of the remaining "BWallocated" bandwidth,

If the used bandwidth by DC $\geq$ 70%

Then the class will be classified in the CE category and sustain the treatment of the Resources Adapter module.

End

The figure 4 illustrates a synthesis of the resources' release algorithm

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

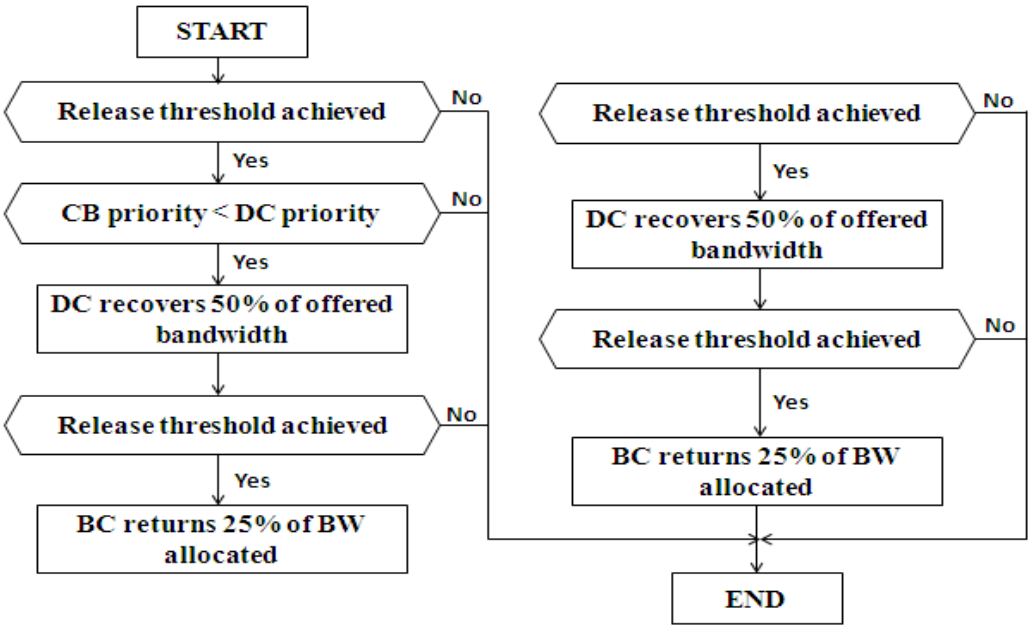


Fig. 4 Resources release diagram

C. *Simulation and evaluation of the PB SAQOS model:*

A. *Simulation model:*

For the evaluation of BP SAQOS model we have installed the topology presented by the [Fig.5], clients sites are connected by the Generic Routing Encapsulation (GRE) tunnels, each client is connected to the provider edge "PE" by E1 connection, the simulated protocols are from different natures: Bulk applications (FTP and TFTP), Signaling application (SCCP), Real time application (RTP) and Best Effort applications (http). For the VOIP simulation, we have installed the Cisco's Call Manager Express technology [21], telephones download their Firmware from the CME by using the TFTP protocol, used Firmware in simulation is 7921 with a size of 10.3MB.

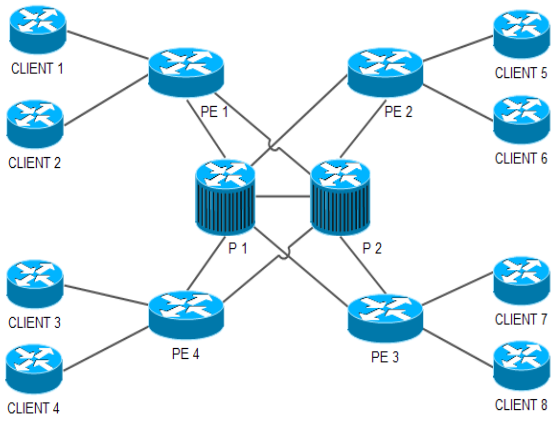


Fig. 5 Experimentation Scale model:

14 VOIP clients are deployed on each client site, with two simultaneous calls to each site, used codec is the G711 64kbps, the average bandwidth required for 14 simultaneous calls is 1176K [22], bandwidth's reservation and DSCP code assignments are performed based on a real example of a communicating corporate network [23,24].

Traffic	Reservation	DSCP	Tool
Voice Signaling	Priority 20%	AF31	SCCP Protocol
Voice Traffic	Priority 40%	EF	Call Manager Express
FTP	Bandwidth 25%	AF11	FileZilla Server
TFTP	Bandwidth 10%	AF11	TFTPD 32
HTTP	Bandwidth 5%	BE	WAMP Server

Table 1 Quality of Service requirements

In order to validate the PB-SAQOS model, we have proceeded by three simulations: without Quality of Service, using Class-Based Weighted Fair Queuing (CBWFQ) and using PB-SAQOS model.

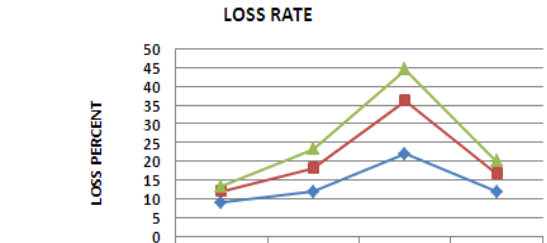
B. *Obtained results:*

The evaluation criteria are:

- Loss rate by flow, the jitter and the latency of the VOIP.

The figure 6, illustrates the loss rate by application, the results obtained without QoS and with CBWFQ are consistent with expectations [25, 26]. PB SAQOS has proven its efficiency, due to the intelligent share of bandwidths between multiple categories.

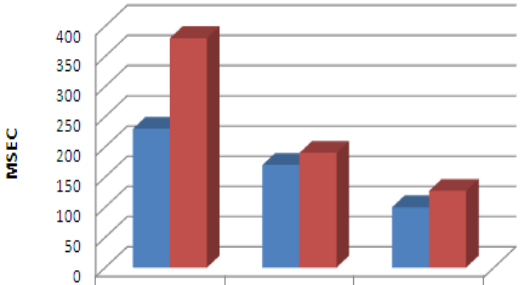
http://sites.google.com/site/ijcsis/
ISSN 1947-5500



	VOIP	FTP	TFTP	HTTP
PB SAQOS	1.2	5	8.3	3.3
CBWFQ	3.1	6.2	14.2	4.8
WITHOUT QOS	9	12	22	12

Fig. 6 Loss rates comparison

The figure 7 represents values of the jitter and latency of the VOIP traffic.



	WITHOUT QOS	CBWFQ	BP SAQOS
JITTER	230	170	100
LATENCY	380	190	127

Fig. 7 Comparisons of the Jitter and the Latency

D. Conclusion :

This paper discussed a novel model of the policy-based smart adaptive quality of service management, the model proposes a distributed data, control and management plans, and new algorithms of resources adaptation and release, for a rapid and efficient QoS adaptation, the model has shown its flexibility and efficiency by proposing best results for different types of traffic.

REFERENCES

[1] Zhang, Q. T., Chen, J. I. A. Y. I., & Zhu, H. O. N. G. B. O. (2014). Network convergence: theory, architectures, and applications. Wireless Communications, IEEE, 21(6), 48-53.

[2] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998). An architecture for differentiated services.

[3] Jacobson, V., Nichols, K., & Poduri, K. (1999). An expedited forwarding PHB.

[4] Heinanen, J., Baker, F., Weiss, W., & Wroclawski, J. (1999). Assured forwarding PHB group (Vol. 470, pp. 471-472). RFC 2597, june.

[5] Bellagamba, E., Ward, D., Mirsky, G., Andersson, L., Takacs, A., & Skoldstrom, P. (2014). Configuration of Pro-Active Operations, Administration, and Maintenance (OAM) Functions for MPLS-based Transport Networks using RSVP-TE.

[6] Shenker, S.—Partridge, C.—Guerin, R.: Specification of Guaranteed Quality of Service. Request for Comments (Standard Track) RFC 2212, Internet Engineering Task Force, September 1997.

[7] Wroclawski, J.: Specification of the Controlled-Load Network Element Service. Request for Comments (Standard Track) RFC 2211, Internet Engineering Task Force, September 1997

[8] Cabuk, S., Dalton, C. I., Eriksson, K., Kuhlmann, D., Ramasamy, H. V., Ramunno, G., ... & Stüble, C. (2010). Towards automated security policy enforcement in multi-tenant virtual data centers. Journal of Computer Security, 18(1), 89-121.

[9] Yoshihara, K., Isomura, M., & Horiuchi, H. (2001). Distributed policy-based management enabling policy adaptation on monitoring using active network technology.

[10] Samaan, N., & Karmouch, A. (2005). An automated policy-based management framework for differentiated communication systems. Selected Areas in Communications, IEEE Journal on, 23(12), 2236-2247.

[11] Shanbhag, S., & Wolf, T. (2011). Automated composition of data-path functionality in the future internet. Network, IEEE, 25(6), 8-14.

[12] Bari, M. F., Chowdhury, S. R., Ahmed, R., & Boutaba, R. (2013, November). PolicyCop: an autonomic QoS policy enforcement framework for software defined networks. In Future Networks and Services (SDN4FNS), 2013 IEEE SDN for (pp. 1-7). IEEE.

[13] Manzalini, A., Saracco, R., Buyukkoc, C., Chemouil, P., Kuklinski, S., Gladisch, A., ... & Mueller, J. (2014). Software-Defined Networks for Future Networks and Services: Main Technical Challenges and Business Implications.

[14] Wang, W., Qi, Q., Gong, X., Hu, Y., & Que, X. (2014). Autonomic QoS Management Mechanism in Software Defined Network. Communications, China, 11(7), 13-23.

[15] Bari, M. F., Roy, A. R., Chowdhury, S. R., Zhang, Q., Zhani, M. F., Ahmed, R., & Boutaba, R. (2013, October). Dynamic controller provisioning in software defined networks. In 2013 9th International Conference on Network and Service Management (CNSM) (pp. 18-25). IEEE.

[16] Koponen, T., Casado, M., Gude, N., & Stribling, J. (2014). U.S. Patent No. 8,830,823. Washington, DC: U.S. Patent and Trademark Office.

[17] Hassas Yeganeh, S., & Ganjali, Y. (2012, August). Kandoo: a framework for efficient and scalable offloading of control applications. In Proceedings of the first workshop on Hot topics in software defined networks (pp. 19-24). ACM.

[18] Aagesen, F. A., & Thongtra, P. (2012, May). On adaptability issues of networked service systems. In Digital Information and Communication Technology and it's Applications (DICTAP), 2012 Second International Conference on (pp. 277-282). IEEE.

[19] Pitt, E., & McNiff, K. (2001). Java. rmi: The Remote Method Invocation Guide. Addison-Wesley Longman Publishing Co., Inc..

[20] Basanta-Val, P., & Anderson, J. S. (2012). Using real-time java in distributed systems: Problems and solutions. In Distributed, Embedded and Real-time Java Systems (pp. 23-44). Springer US.

[21] Cisco IP Communications Express: CallManager Express with Cisco Unity Express. Cisco Press, 2005.

[22] Press, Cisco. "Voice over IP-Per Call Bandwidth Consumption." (2005).

[23] Babiarz, J., Chan, K., & Baker, F. (2006). Configuration guidelines for DiffServ service classes. IETF-Request for Comments (RFC 4594), (4594).

[24] Floyd, S., & Jacobson, V. (1995). Link-sharing and resource management models for packet networks. Networking, IEEE/ACM Transactions on, 3(4), 365-386.

[25] Liu, X., & Tu, C. (2011). An VoIP Application Design with Dynamic QoS Control. In Theoretical and Mathematical Foundations of Computer Science (pp. 605-612). Springer Berlin Heidelberg.

[26] Szilágyi, S. (2013). Analysis of the algorithms for congestion management in computer networks. Carpathian Journal of Electronic & Computer Engineering, 6(1).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Low footprint Hybrid Finite field multiplier for Embedded cryptography

Sunil D. Bobade
Research Scholar,
S.G.B.Amravati University
Amravati, India

Dr.Vijay R. Mankar
Deputy Secretary,RBTE,
Pune Region,
Pune,India

***Abstract*—Finite field multiplier contributes significantly to the area occupancy of cryptoprocessor. In this paper, we propose a novel hybrid finite field multiplier that invokes the more efficient multiplication algorithm. The proposed multiplier switches between two variants of multiplier depending on the size of multiplicands. The Karatsuba multiplier is efficient algorithm ensuring fewer LUTs and stable number of Flip-flops for the smaller bit multiplications, while the other systolic variant ensures fewer LUTs count for the bigger size multiplicands. The proposed hybrid multiplier does the initial recursion using the systolic variant while final small sized multiplications are accomplished using the Karatsuba algorithm. Area analysis report suggests that using a proposed hybrid multiplier instead of just traditional Karatsuba Multiplier, eventually helps in reducing FPGA footprint.**

Keywords-Karatsuba Multiplier; Modular Multiplication; Cryptoprocessor; Area complexity.

I. INTRODUCTION

The need for security in embedded applications is the most ignored design practice due to resource constraints. Implementing security protocols for embedded systems are often restricted due to complexity, limited resources, memory inadequacy, processor incapability, energy and area constrain hence embedded designers have paid a little heed to secure embedded system. As a result lower levels of security were installed to protect embedded systems. Very recently, system designers felt the need to secure embedded systems as the attack vulnerability grew and network related applications started using embedded systems. Cryptography techniques tailor made for embedded systems started pouring in with a major thrust on space optimization. The Area occupancy of most popular public key algorithm is based on modular arithmetic, where the most critical operation is modular multiplication [1]. Hence by targeting area occupancy of modular multiplier, the overall area complexity of security scheme can be drastically reduced.

The design of Finite field multiplier in crypto processor is the most critical design issue in the implementation of the ECC processor as it occupies the major portion of Cryptoprocessor. A number of multipliers with different area complexity are reported in the available literatures. The Karatsuba algorithm is agreed upon as a most efficient multiplication algorithm and is widely adopted in VLSI implementation of cryptoprocessor. We have analyzed the area impact of integrating traditional Karatsuba multiplier with a systolic multiplier. Instead of using traditional Karatsuba Multiplier alone, we have integrated it with finite field multiplier which adopts a systolic approach. The space complexity of the resulting hybrid multiplier is found to be much better than those of traditional Karatsuba multiplier. This is a significant achievement if we intend to use this multiplier as a part of crypto processor in embedded systems for performing modular multiplication activities.

II. RELATED WORK

Several area and speed optimized multipliers have been proposed and developed to cater critical multiplication activity in cryptography in [2,3]. Of them all, Karatsuba- Ofman algorithm [4] is widely accepted to be highly area and speed optimized. Due to its simplicity and efficiency, its polynomial version is widely followed in VLSI implementations of parallel multipliers in $GF(2^n)$ based cryptosystems. In [2,4,] a variant of Karatsuba multiplier of the type $GF((2^n)^8$ is presented. Ashkan Hosseinzadeh Namin, Huapeng Wu,and Majid Ahmadi proposed a word-level modular multiplier using normal basis [5]. In [6] Hossein Mahdizadeh and Massoud Masoumi achieved efficiency in multiplication by placing multipliers in parallel. Bit-parallel [7], bit-serial [8], digit-serial/parallel [9], multipliers or systolic architectures [10] are the few of the existing solutions already explored. Each type of these solutions has different pros and cons: bit-serial are slow but small, bit-parallel solutions are fast but larger. Digit as well as systolic architectures provides some tradeoff between speed and area occupancy.

Elliptic curve scalar multiplier with improved area and speed was designed by Sujoy Sinha Roy *et.al* [11]. Point addition and doubling operations were intelligently scheduled. Multiplier had a three stage pipelined architecture for double and add based scalar. The implementation used a novel pipelined bit-parallel Karatsuba multiplier with subquadratic complexity. In this design efficient choice of scalar multiplication algorithm and enhanced scheduling of point arithmetic contributed in a high-speed architecture with a significantly small area.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Hossein Mahdizadeh and Massoud Masoumi [12] designed architecture for efficient FPGA implementation of elliptic curve cryptographic processor over GF (2^{163}). In the implemented design the execution delay of the Lopez-Dahab scalar point multiplication algorithm has been optimized by parallelization of the multipliers.

A tripartite modular multiplication was proposed by Kazuo Sakiyama et.al [13]. To achieve efficiency this algorithm effectively utilizes and merges three different existing algorithms, classical modular multiplication based on Barrett reduction, the modular multiplication with Montgomery reduction and the Karatsuba multiplication algorithms. In multiprocessor environment for hardware and software implementations, this algorithm is very efficient. This modular multiplier clocks higher speed compared to the other for modular multiplication algorithms.

The Massey-Omura multiplier uses normal basis representations of the field elements. This representation helps in making the structure of the multiplier highly uniform resulting in efficient hardware architecture. The architecture processes parallel input but the result is produced serially [14]. Another efficient multiplier based on normal basis is the *Sunar-Koç* [15] multiplier. The multiplier is area efficient compared to the Massey-Omura multiplier and has similar timing requirements.

III. PROPOSED SCHEME

Here we propose two versions of finite field multiplier and will systematically and intelligently switch between two versions. Two versions use different algorithms for multiplication and each version has its own advantage. By clubbing the two multipliers, area investigation suggests area optimization is possible. The first Multiplier variant is the traditional basic recursive Karatsuba multiplier.

FFmul *(FF Efficient Multiplier Algorithm)*

Input: A,B are multiplicands of m bits
Output: C of length 2m − 1 bits
/* **Define :** Mx → AxBx */
/* **Define :** M(x,y) → (Ax + Ay)(Bx + By) */
begin
for i = 0 *to* m − 2 **do**
Ci = C_{2m-2-i} = 0
for j = 0 to[i/2] **do**
if i = 2j **then**
$C_i = C_i + M_j$
$C_{2m-2-i} = C_{2m-2-i} + M_{m-1-j}$
else
$C_i = C_i + M_j + M_{i-j} + M_{(j,i-j)}$
$C_{2m-2-i} = C_{2m-2-i} + M_{m-1-j} + M_{m-1-i+j} + M_{(m-1-j,m-1-i+j)}$
end
end
$C_{m-1} = 0$
for j = 0 to [(m − 1)/2] **do**
if m − 1 = 2j **then**
$C_{m-1} = C_{m-1} + M_j$
else
$C_{m-1} = C_{m-1} + M_j + M_{m-1-j} + M_{(j,m-1-j)}$
end
end
end

Upon investigation, up to the multiplicand size 29, this algorithm uses fairly stable number of flipflops but LUTS rises phenomenally as multiplicand size rises above 29. Maximum number of LUTs required for performing 256 bit multiplication is high as 1535.Thus; the FF efficient multiplier version alone is not efficient for FPGA platforms because of large number of LUTs involved resulting in larger footprints. Hence up to the multiplicand size 29, this algorithm is most suited.

Second version of the multiplier, splits word into least significant and most significant words and using the shifting operation generates the product. This multiplier algorithm is found to be highly LUT efficient after m>29

LUTmul (LUT Efficient Multiplier algorithm)

Input: The multiplicands A, B and their length m
Output: C of length 2m − 1 bits
begin
l = [m/2]
A′ = A[m−1….l] + A[l−1….0]
B′ = B[m−1…..l] + B[l−1….0]
C_{p1} = hmul(A[l−1…0],B[l−1…0], l)
C_{p2} = hmul(A′ ,B′ , l)
C_{p3} = hmul(A[m−1…l],B[m−1…l], m − l)
return (C_{p3} left shift 2l) + (C_{p1} + C_{p2} + C_{p3}) left shift l + C_{p1}
end

After having two versions, FF efficient multiplier working well for < 29 and another version, LUT efficient multiplier for m> 29, Hybrid Multiplier intelligently switches between two versions on the threshold of 29.

In proposed hybrid Finite Field multiplier based on Hmul Algorithm, the m bit multiplicands are split into two parts. When the number of bits is greater than or equal to the threshold 29, LUT efficient version is invoked. As the number of bits of the multiplicand falls less than 29, the FF efficient variant of Karatsuba algorithm is invoked. The proposed hybrid multiplier does the initial recursion using the LUT efficient variant algorithm while final small sized multiplications are accomplished using the Karatsuba algorithm.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Hmul Hybrid Multiplier algorithm

Input: The multiplicands A, B and their length m
Output: C of length 2m − 1 bits
begin
  if m < 29 **then**
    return FFmul(A,B,m)
  else
    return LUTmul(A,B,m)
  end
end

IV. Results and Discussion:

In this section, we focus on the FPGA implementation of the proposed hybrid multiplier. The proposed architecture is coded in verilog HDL and is synthesized using Xilinx ISE version 14.4 design software and is implemented on Xilinx Virtex-4 xc4vlx200ff1513 FPGA. The RTL schematic and the simulation result for the Hybrid Finite Field multiplier is shown in Fig. 1 and Fig. 2.

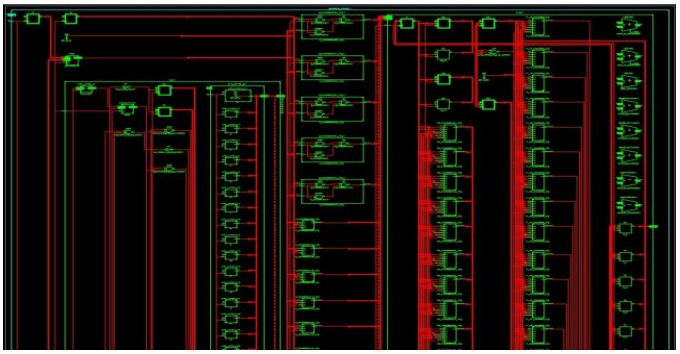


Figure 1 RTL Schematic for Hybrid finite field Multiplier

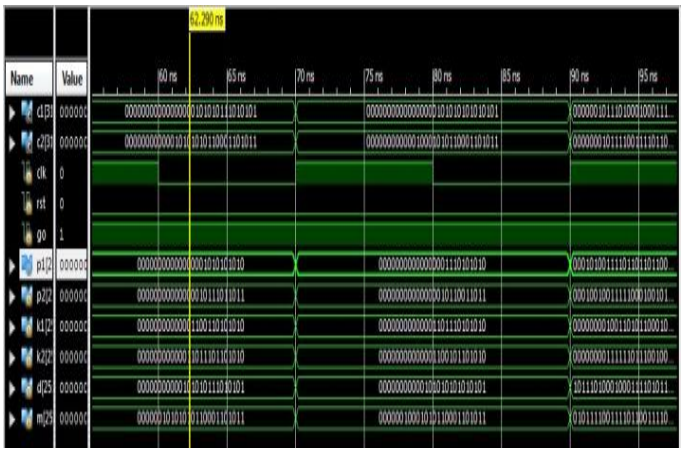


Figure 2 Simulation results for Hybrid Multiplier

A. A rea Report of Hybrid Multiplier:

Since the area of the complete processor mainly depends on the incorporated GF multiplier, most of the slices in the target device are utilized by it. From table, it is observed FF efficient multiplier involves fewer and stable number of flipflops for m<29 and the other version uses almost stable number of flipflops and fewer LUTs for m>29 as shown in table 1.

Hence, our hybrid multiplier intelligently switches between two versions depending on the size of m and makes optimum use of resources available.

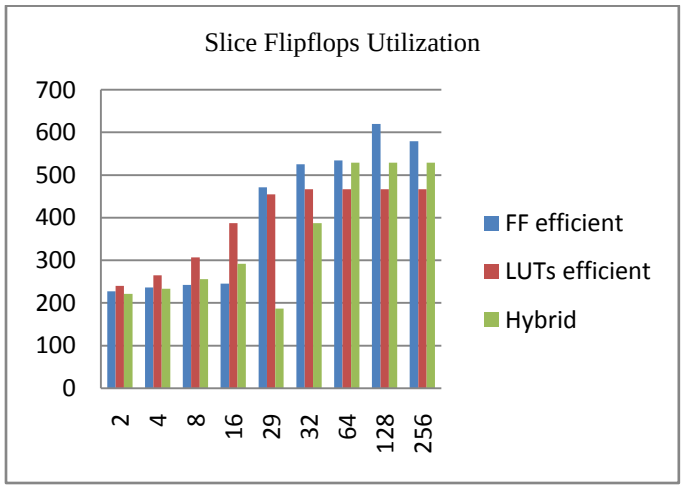


Fig. 3 Flipflops Utilization by the three multipliers

Fig. 3 indicates for m< 29, almost stable number of Flipflops is used by FF efficient Multiplier. But for higher values of m beyond 29, this version uses higher resources to accomplish multiplication task.

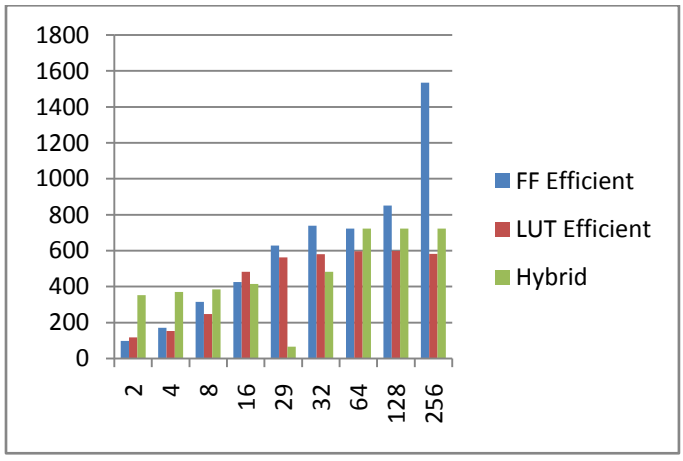


Fig. 4 LUTs Utilization by the three multipliers

Fig. 4 suggests for m> 29, LUT efficient version utilizes almost stable and fewer number of LUTs , resulting in a lower footprint.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1: Device utilization summary of three Multipliers

	FF Efficient Multiplier				LUT efficient Multiplier				Hybrid Multiplier			
Multiplicand sizes	*Slice Flipflops*	*LUTs*	*Slices*	*IOBs*	*Slice Flipflops*	*LUTs*	*Slices*	*IOBs*	*Slice Flipflops*	*LUTs*	*Slices*	*IOBs*
2	227	98	124	259	240	118	127	259	221	353	217	199
4	236	171	168	259	265	153	146	259	233	371	230	203
8	242	316	237	259	307	248	191	259	256	384	241	211
16	245	426	292	259	387	482	307	259	292	414	267	227
29	471	629	412	259	455	562	362	259	187	66	97	253
32	525	739	473	259	467	580	374	259	387	482	307	259
64	534	723	469	259	467	597	381	259	529	723	480	323
128	620	851	576	259	467	599	382	259	529	723	480	323
256	579	1535	916	259	467	583	374	259	529	723	480	323

The resultant hybrid multiplier uses fairly stable number of flipflops as multiplicand size is varied up to 29 and fewer and stable number of LUTs as the multiplicand size is varied beyond 29. Hybrid multiplier performs higher bit multiplication using LUT efficient version algorithm and switches to Karatsuba which is not a LUT efficient multiplier as the multiplicand size falls below 29. Hybrid multiplier uses maximum of 292 slice flipflops, 414 LUTs and 267 slices in contrast to traditional Karatsuba multiplier which needs 245 slice flipflops, 426 LUTs and 292 slices for a typical 16 bit multiplication activity.

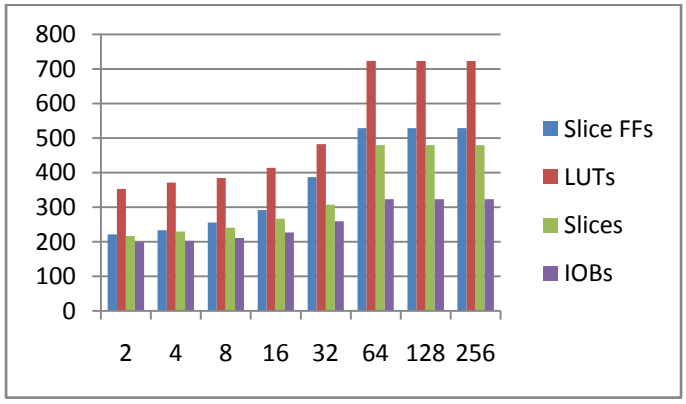


Fig. 5 Device Utilization Summary of Hybrid multiplier for different multiplicand sizes

Fig. 5 shows the bar chart representation of the resource utilized by the proposed Hybrid Multiplier with respect to the Operand and Size digit size of the multiplier. With the increase in Operand size or digit size for multiplier design, the number of FFS, LUTs and slices remains fairly stable.

A bar chart representation shown in Fig. 6 compares the resource utilization of Proposed Hybrid multiplier with traditional Karatsuba Multiplier for m=256. Thus proposed hybrid multiplier exhibits a savior of 7.56 % in terms of Flip flop slices. Proposed Multiplier involves 52 % fewer LUTs and utilizes 47% fewer slices as compared to traditional Karatsuba multiplier for 256 bit multiplication. This helps in bringing down the footprint of modular multiplier on FPGA while building cryptography schemes.

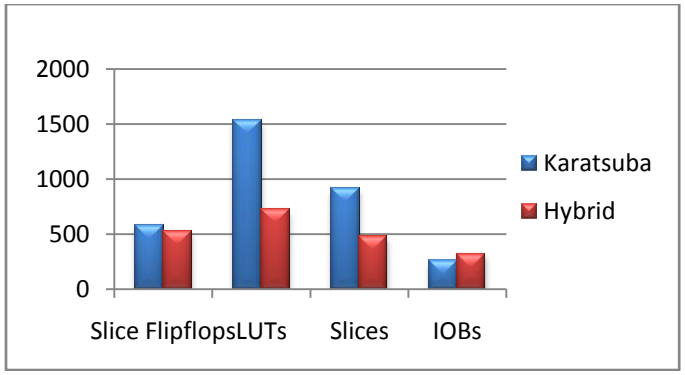


Fig. 6 Device Utilization Comparisons of Karatsuba and Hybrid multiplier for m=256

V. CONCLUSION

A novel method has been proposed to implement the finite field multiplier. Area analysis report suggests that using a proposed hybrid multiplier instead of just traditional Karatsuba Multiplier, eventually helps in reducing FPGA footprint. This is a significant achievement if we intend to use this multiplier in VLSI implementation for performing modular multiplication arithmetic in embedded cryptography.

REFERENCES

[1] Blake, I. F., Seroussi, G., and N. P. *Elliptic curves in cryptography*. Cambridge University Press, New York, NY, USA, 1999.
[2] C. Paar. *Efficient VLSI Architectures for Bit Parallel Computation in Galois Fields*. PhD thesis, University at GH Essen, VDI Verlag, 1994.
[3] B. Sunar and C¸ . K. Koc¸. Mastrovito multiplier for all trinomials. *IEEE Transactions on Computers*, 48(5):522–527, May 1999.
[4] C. Paar. A new architecture for a paralel finite field multiplier with low complexity based on composite fields. *IEEE Transactions on Computers*, 45(7):856– 861, July 1996.

[5] Ashkan Hosseinzadeh Namin, Huapeng Wu,and Majid Ahmadi, "A word-level finite field multiplier using normal basis", IEEE Transactions on computers, Vol. 60, No. 6, pp: 890- 895, Jun. 2011
[6] Hossein Mahdizadeh and Massoud Masoumi, ""Novel architecture for efficient FPGA implementation of elliptic curve cryptographic processor over GF(2^{163})", IEEE Transactions on very large scale integration (vlsi) systems, Vol. 21, NO. 12, pp: 2330- 2333, Dec.2013
[7] Y. I. Cho, N. S. Chang, C. H. Kim, Y.-H. Park, and S. Hong, "New Bit Parallel Multiplier With Low Space Complexity for All Irreducible Trinomials Over," IEEE Trans. VLSI Syst., vol. 20, no. 10, pp. 1903–1908, Oct. 2012.
[8] M. Morales-Sandoval, C. Feregrino-Uribe, and P. Kitsos, "Bit-serial and digit-serial GF (2m) Montgomery multipliers using linear feedbackshift registers," Computers Digital Techniques, IET, vol. 5, no. 2, pp.86–94, Mar. 2011.
[9] A. Hariri and A. Reyhani-Masoleh, "Digit-Serial Structures for the Shifted Polynomial Basis Multiplication over Binary Extension Fields," WAIFI 2008, LNCS 5130. Springer, pp. 103–116, Jul. 2008.
[10]J. Lin, "Low-latency Digit-serial Systolic Double Basis Multiplier over GF (2 m) using Subquadratic Toeplitz Matrix-vector Product Approach," IEEE Trans. Comput., vol. PP, no. 99, p. 1, 2012
[11] Roy. S.S, Rebeiro,C and Mukhopadhyay. D, "Theoretical modeling of elliptic curve scalar multiplier on LUT-based FPGAs for area and speed", IEEE Transactions on Very Large Scale Integration (VLSI) systems, Vol. 21, No. 5, May 2013.
[12] Hossein Mahdizadeh and Massoud Masoumi, "A novel architecture for efficient FPGA implementation of elliptic curve cryptographic processor over GF(2^{163})",IEEE Transactions on very large scale integration systems, Vol. 21, No. 12, Dec 2013
[13] Kazuo Sakiyama, Miroslav Knezevica, Junfeng Fana, , Bart Preneela, and Ingrid Verbauwhedea, "Tripartite modular multiplication", Integration, the VLSI Journal, Vol. 44, No.4, pp: 259–269, September 2011.
[14] Gregory C. Ahlquist, Brent E. Nelson, and Michael Rice, "Optimal Finite Field Multipliers for FPGAs," in *FPL '99: Proceedings of the 9th International Workshop on Field-Programmable Logic and Applications*, London, UK, 1999, pp. 51–60, Springer-Verlag.
[15] Ç. K. Koç and B. Sunar, "An Efficient Optimal Normal Basis Type II Multiplier," *IEEE Trans. Comput.*, vol. 50, no. 1, pp. 83–87, 2001.

AUTHORS PROFILE

Sunil D. Bobade obtained his Engineering Graduate Degree in Electronics and Telecommunication Engineering from VYWS College of Engineering, Amravati (India) in 1994, Post Graduate Degree in Electronics Engineering from S.G.B.Amravati University, Amravati (India) in 2002.

He has been in the field of teaching since last 19 years and is presently, working as an Assistant Professor in Department of Information Technology, Datta Meghe College of Engineering, Navi Mumbai. He is also working as a research scholar in S.G.B.Amravati University and is working on development of area efficient algorithms for protection of memory of embedded systems.

Dr.Prof.Vijay R. Mankar received his B.E. degree in Electronics and Power Engineering from Government College of Engineering, Amravati (India) in 1986, M.Tech. in Electronics Engineering from erstwhile Visvesvaraya Regional College of Engineering, Nagpur (India) in 1990 and Ph.D. from S.G.B. Amravati University, Amravati (India) in 2009.

He has been in the field of teaching since last 23 years and is presently, working as a Deputy Secretary, MSBTE, Pune Region. He has served as Head Of Department and the Professor, with Department of Electronics Engineering, Govt. Polytechnic, Amravati. He has been active in the research as well. His research interests are in Neural Networks, Design of Embedded system, Data security. He has published 19 research papers in journals and conferences of national and international repute.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Base Station Radiation's Optimization using Two Phase Shifting Dipoles

Safaa BERRA,
Laboratory RITM/ESTC,
ENSEM,University of Hassan II,
Casablanca, Morocco

Mounir Rifi
Laboratory RITM/ESTC,
ENSEMUniversity of Hassan II,
Casablanca, Morocco

***Abstract*—The electromagnetic pollution becomes a serious issue, with the increasing number of telecom operators by country. Certainly the technologies differ, in term of naming (BTS, NodeB, E-NodeB), in term of frequencies, but the problem remains the same: how to optimize the radiation of base station? Is there any intelligent system which allows us to reduce the radiation area and to channel it according to the demand? Far from the complicated and expensive networks of antennas, we are going to present in this paper the use of a simple system of antennas to optimize the base station's radiation.**

We will start this article by presenting some techniques that improve the energy consumption in base station which represent an important ratio of the global energy in a cellular network. Then, we will study the phase shifting effect on the gain and the energy efficiency. This paper aims to use the found results to propose a new process to reduce the radiation of base station.

In the third part, we propose a new architecture based on the use of two phased dipoles in the base station to cover in a smart way the subscribers as long as they are present in the cell. We propose a new system that allows us to reduce the radiation area.

Keywords*—Smart antenna, Phased array system, Dipole, Base station, Energy consumption, Energy efficiency, Gain, Radiation.

I. INTRODUCTION

The scientific community considers that the only known sanitary effects of the radio frequencies are thermal short-term effects. However, researches and experiences led on several people showed a relation between the excessive exposure to the electromagnetic field and the appearance of health problems [1] but no demonstration was brought in this sense [2].

The exposure to magnetic fields is regulated at the international level by restrictions [3] and reference levels [4], see TABLE I and TABLE II. These recommendations aim to provide a high level of protection of the public against the potentially harmful effect.
It is one of the most challenging issues to manage the base station's radiation in a way to optimize and reduce the radiation area and at the same time to preserve the quality of service required for mobile communication. We need to have a heavy, simple system that we can implement easily in the existent base stations.

TABLE I.
THE BASIC RESTRICTIONS OF THE FREQUENCIES BETWEEN 10 AND 300GHZ

Type of exposure	Power density (W/m²)
Occupationnal exposure	50
General public exposure	10

TABLE II.
THE REFERENCE LEVELS AT 900MHZ AND 1800 MHZ

Frequency	Intensity of the electric field	Intensity of the magnetic field	Power density
900MHz	41 V/m	0,1 A/m	4,5 W/m²
1800 MHz	58 V/m	0,15 A/m	9 W/m²

This paper is organized as follows. First, we will study the evolution of the gain and the energy efficiency after adding the phase shifting, according to the number of dipoles.
The objective is to capitalize on the existing intelligent systems of antennas such as adaptive array system and switched beam system, to benefit from the advantages of both methods and remedy to their deficiencies [5].

In the second part we enumerate the main components of the new architecture that we will propose, and we describe the process and the interoperability allowing these components to perform the radiation command according to the phase shifting. Scenarios of covering subscribers in the base station area will be illustrated and explained in detail. In addition to that, some necessary conditions will be modeled as an inequation.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. IMPROVEMENT OF THE ENERGY EFFICIENCY OF BASE STATIONS BY THE USE OF THE PHASED ARRAY ANTENNA

A. Energy Consumption in Base Station

A mobile cellular network consists of three key elements; a core network which insures the switching, a base station acting as radio interface, in addition to the mobile terminals.
All the components of the cellular network consume some energy. However, the major part of the energy consumption comes from the network of radio access, indeed and as we can notice it on Figure 1 base stations are consuming around 57 % of the energy of a typical cellular network [6].

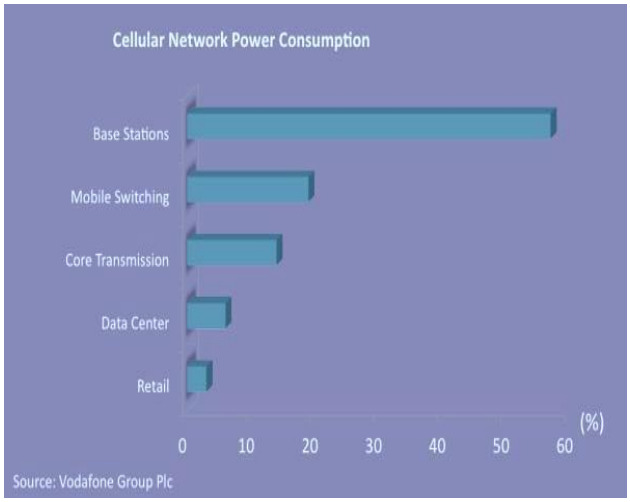


Figure 1. Percentage of energy consumption in the infrastructures of a typical cellular network

Base stations represent up to 80 % of the energy consumed on the whole network for some operators [6].
Some techniques are already used by providers and mobile operators to improve the energy efficiency of base station:

- **Eco-energetic power amplifiers:** Usually a power amplifier consumes around 50 % of the energy used by the base station. Approximately 80 to 90 % of this energy is wasted in the form of heat, thus adding another cost of energy due to the use of an air conditioner [6].
- **Base station sleep mode:** The idea is to save energy by switching to sleep mode, the largest number of base stations during the period of low traffic, without reducing the covered area or the availability of the service.
- **The control of power:** Reducing power emission according to the distance between BTS and mobile station.
- **Green source of energy**: Using for example the solar panels instead of electricity for the BTS feed.

In the next paragraph we will study the evolution of the gain and the energy efficiency for an example of directional antenna.

B. Evolution of the Gain by the Introduction of the Phase Shifting

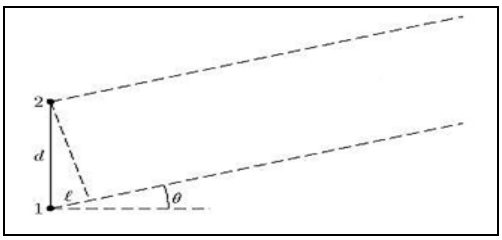


Figure 2. Two dipoles radiating towards a distant point

Let's consider two identical half wave dipoles distant to each other by a distance d. We are going to calculate the electric field produced by this pair of dipoles at a distance r very far from the antennas.
As we have r>> d, the angle Θ is considered the same for both dipoles and the produced field will also be the same. The dipole n° 2 is closer to the point of radiation than the dipole n°1 as illustrated in Figure 2. The field produced by the dipole n°2 will arrive earlier than the field produces by the dipole n°1. In other words, the field produced by the dipole n°2 will have a phase advance of k*d*sin(Θ). Taking into account the supplementary phase difference between currents φ the final expression of total phase shifting [7] is:

$$\psi = k*d*\sin(\theta) - \varphi \quad (1)$$

Where $k = \frac{2*\pi}{\lambda}$, $d = \frac{\lambda}{2}$. λ is the wavelength and φ is the phase difference between currents.
As both fields are parallel, we will consider the following sum:

$$|E_\theta| = 2*|E_{\theta 1}|*\cos(\psi/2) \quad (2)$$

$$|E_\theta| = 2*|E_{\theta 1}|*\cos(k*d*\sin(\theta)/2 - \varphi/2) \quad (3)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

To maximize the value of both the field and the characteristic function of radiation we have this relation between phase and angle :

$$\varphi = k*d*\sin(\theta) \quad (4)$$

Let us take a subscriber being in the direction of $\pi/6$. From the relation (4) to have the radiation towards this angle it is necessary to introduce a phase shifting of $\pi/2$ in the system of two phased dipoles. We assume that there is no loss and that the gain is equal to the directivity. The expression of the gain is:

$$G(\theta) = 4*\pi * \frac{F^2(\theta)}{\oint_\Omega F^2(\theta,\phi)d\Omega}\,[\mathrm{dBi}] \quad (5)$$

Where dΩ is the solid angle and F is the characteristic function of radiation.

For two phased dipoles:

$$F_{\text{two phased dipoles}}(\theta) = \cos(\pi/2*\sin(\theta) - \varphi/2) \quad (6)$$

$$G_{\text{two phased dipoles}}\left(\frac{\pi}{6}\right) = 0.96\ \mathrm{dBi} \quad (7)$$

For two simple dipoles without phase shifting:

$$F_{\text{two simple dipoles}}(\theta) = \cos(\pi/2*\sin(\theta)) \quad (8)$$

$$G_{\text{two simple dipoles}}\left(\frac{\pi}{6}\right) = 0.62[\mathrm{dBi}] \quad (9)$$

We can notice that the phase difference between currents allows a gain of 50% in the system of two dipoles.

C. Evolution of the Gain According to the Number of Dipoles

In the conditions of far fields, the characteristic function of radiation for one dipole is:

$$F_{\text{one dipole}}(\theta) = \sin(\theta) \quad (10)$$

The gain obtained with one simple dipole is:

$$G_{\text{one dipole}}\left(\frac{\pi}{6}\right) = 0.31[\mathrm{dBi}] \quad (11)$$

The value of the gain obtained in the direction of $\pi/6$ with two phased dipoles is three times important compared to a simple dipole.

If the condition (4) between phase and angle is true, the characteristic function will be maximized and equal to 1.Thus the results obtained for the angle $\pi/6$ can be generalized for any direction in the space.

Figure 3 illustrates the evolution of the gain according to the number of dipoles in the uniform linear phased network.

$$G_{\text{three phased dipole}}\left(\frac{\pi}{6}\right) = 1.73[\mathrm{dBi}] \quad (12)$$

$$G_{\text{seven phased dipole}}\left(\frac{\pi}{6}\right) = 6.67[\mathrm{dBi}] \quad (13)$$

We notice that with seven dipoles the value of the gain becomes very important.

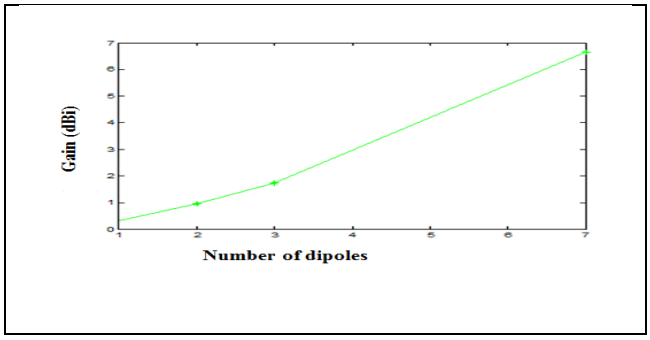


Figure 3. Evolution of the gain according to the number of dipoles

D. Evolution of the Energy Efficiency According to the Number of Dipoles

In the hypothesis of far field, we assume that we have the same phase difference for the electric and magnetic fields. The expression of electromagnetic fields for a simple dipole is:

$$\vec{B} = -\frac{1}{r*c} * \left[\frac{\mu*I*\omega}{4*\pi}\right] * \sin(\theta) * e^{j(\omega t - kr)}\,\vec{u}\varphi$$

$$\vec{E} = -\frac{1}{r} * \left[\frac{\mu*I*\omega}{4*\pi}\right] * \sin(\theta) * e^{j(\omega t - kr)}\,\vec{u}\theta \quad (14)$$

Where r is the distance to the far point, c is the speed of propagation of the light, μ is the permeability, I is the intensity of current, and $\omega = 2*\pi*f$, $k = \frac{2*\pi}{\lambda}$, where λ is the wavelength, f is the frequency.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The power density for one dipole is:

$$\vec{E} \wedge \vec{B} = - \alpha * \left[\frac{\sin(\theta)}{r}\right]^2 * e^{2*j(\omega t - kr)} \vec{u}r \quad \textbf{(15)}$$

Where $\alpha = \frac{1}{c} * \left[\frac{\mu * I * \omega}{4 * \pi}\right]^2$

For a uniform linear network of N phased dipoles, the expression of electromagnetic fields, in the hypothesis of far field is:

$$\vec{E}t = N*\vec{E} \; *\cos((k * d * \cos(\theta) - \varphi)/2)$$

$$\vec{B}t = N*\vec{B} \; *\cos((k * d * \cos(\theta) - \varphi)/2) \quad \textbf{(16)}$$

Where d= $\frac{\lambda}{2}$, λ is the wavelength and φ is the phase difference between currents.

For any θ in the space:

$$\frac{dP(\text{N phased dipole})(\theta)}{dP(\text{1 dipole})(\theta)} = N^2 \quad \textbf{(17)}$$

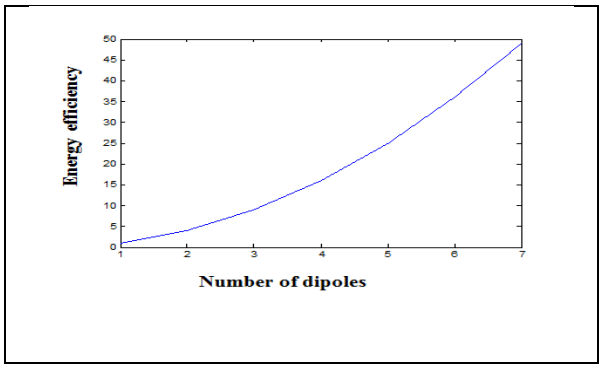


Figure 4. Evolution of the energy efficiency of a uniform linear network of N dipoles compared with only one dipole according to the number N

According to the values obtained in the conditions of far fields, we notice that both gain and energy efficiency increase in the uniform networks linear with numerous dipoles.

III. APPLICATION IN BASE STATIONS

A. *Description of the Components of the Proposed Architecture*

We suggest applying the system of two phase shifting dipoles to a base station. This implementation requires the use of the components below:

- Current
- Two dipoles
- Delay line which includes a predefined table of phase difference values.
- Subscribers' detection system with access to the VLR.
- Microcontroller which allow the control of the interactions between the various components. It will manage the communication with the subscribers' detection system. Besides it defines the table of the corresponding phase difference which will be transmitted to the delay line to perform the beamforming allowing covering the subscribers declared present.

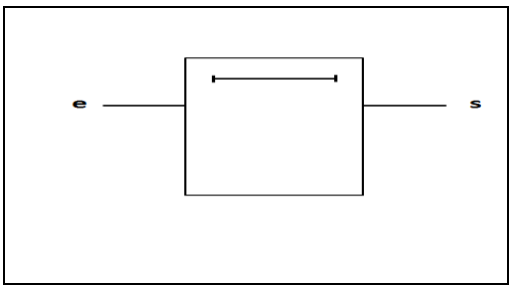


Figure 5. Example of a delay line

$$s(t) = e(t - \tau) \quad \textbf{(18)}$$

Where τ is the delay caused by the line.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table III.
EXAMPLE OF A TABLE OF PHASE DIFFERENCE FOR A TWO PHASED DIPOLE SYSTEM WITH ANGULAR STEP OF 15°

Phase φ	Direction of radiation θ
π ∗ sin(15)	15°
π ∗ sin(30)	30°
π ∗ sin(45)	45°
π ∗ sin(60)	60°
π ∗ sin(75)	75°
π ∗ sin(90)	90°
π ∗ sin(105)	105°
π ∗ sin(120)	120°
π ∗ sin(135)	135°
π ∗ sin(150)	150°
π ∗ sin(165)	165°
π ∗ sin(180)	180°
π ∗ sin(195)	195°
π ∗ sin(210)	210°
π ∗ sin(225)	225°
π ∗ sin(240)	240°
π ∗ sin(255)	255°
π ∗ sin(270)	270°
π ∗ sin(285)	285°
π ∗ sin(300)	300°
π ∗ sin(315)	315°
π ∗ sin(330)	330°
π ∗ sin(345)	345°
π ∗ sin(360)	360°

B. Processes and Operations

Let's consider two identical half wave dipoles distant to each other by a distance d equal to a half wave. A current will be placed in the entry of the first one dipole, and then a phase difference will be introduced on the same current before it arrives at the entry of the second dipole. This phase shifting will be added under the order of the microcontroller via a delay line [8] which we can command afterward.
The command of phase is certainly an important element in this system of antennas, but is not enough, if we want to apply it to GSM base station.
We should remember in the all processing of this new system that the antennas of base station are supposed to cover all the cell area, in the presence or not of subscribers, taking into account the change of spatial coordinate for the subscribers in movement besides the dynamic database of the subscribers.

In parallel with the command of phase, a scanning of the coverage area must be made in a continuous way in a lapse of time unnoticed by the users in mobile communications. This scanning has for objective the localization of every new user who joins or leaves the cell, in an immediate way.

The concerned GSM cell is subdivided into very small areas. Each one of these small areas is covered by the radiation if there is any subscriber's presence.

One or several subscribers will be served by antennas with command of phase.

The phase shifting and the scanning will be seamless for subscribers because the necessary duration for the scan as well as for the beamforming in the desired direction will be in conformity with 3GPP's conditions **[9]**.

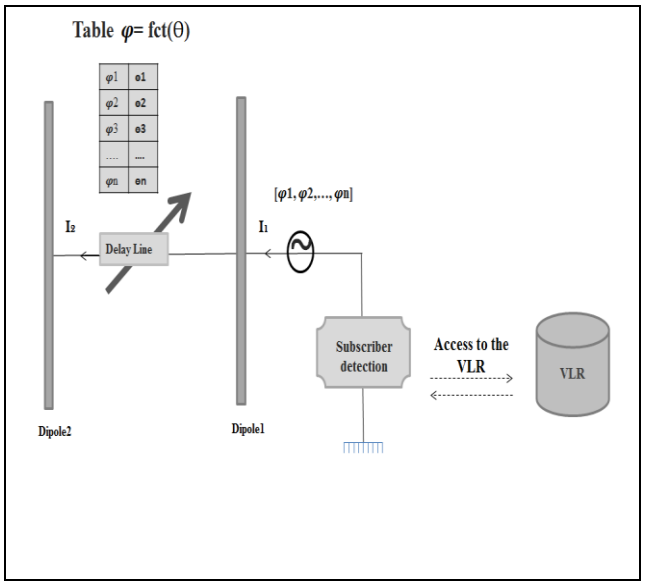


Figure 6. System of two dipoles with command of phase, interaction with the VLR and the delay line

C. Study of the Scenarios of Subscriber's Coverage in a Cell of a Base Station Provided of Two Dipoles with Command of Phase

In the practice, we propose implementing two dipoles with phase shifting in a GSM base station. The phase difference between both currents feeding the two dipoles allows us to command the direction of the main lobe of the resultant radiation pattern. This technique allows reducing the energy lost in the space because we are going to target the subscribers. Besides the obtained gain is much more interesting than the frequent case of simple antenna which shines in all the space**.**

In this paragraph we are going to describe in detail the functioning of the proposed architecture through various possible scenarios. We are going to explain the management of the subscriber's addition and retreat within the same cell. The presence of two or several subscribers in similar, symmetrical and asymmetrical directions. We will define the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

limits conditioning the smooth running of this system to stay in compliance with the 3GPP's standards and avoid the impact on the mobile communications during the various beamforming.
We are going to follow a chronological order throughout paragraphs below to describe the real functioning of base station provided with the system of two dipoles with phase shifting.

1) Case of One Subscriber

At the moment t:

Starting up of the system of two dipoles with phase shifting implemented at the base station. In the practice this moment corresponds to the activation of the BTS

At the moment t+ Δt :

The scanning of all the space during the duration Δt. It allows detecting the number and the location of present subscribers in the cell. It can be also represented by the consultation of the VLR (Visitor Location Register).

At the moment t+Δt+ Δt_1 :

At the moment **t + Δt** and further to the end of the scanning one subscriber is detected in the cell. This subscriber is covered by the radiation. The processing of subscriber's covering, the beamforming in the wished direction has to be made in duration Δt_1 of whose the sum with duration Δt of scanning of the cell is in conformity with the 3GPP's standards, thus not impacting on the mobile communications. We should be in conformity with the condition Δt_1 +Δt <=50ms.
The subscriber's covering consists in the generation of two lobes in the main direction and the corresponding symmetric direction. The generation of both lobes is made at the level of both dipoles via the mechanism of beamforming, realized in our case by the delay line leading the phase difference between the currents of both dipoles.

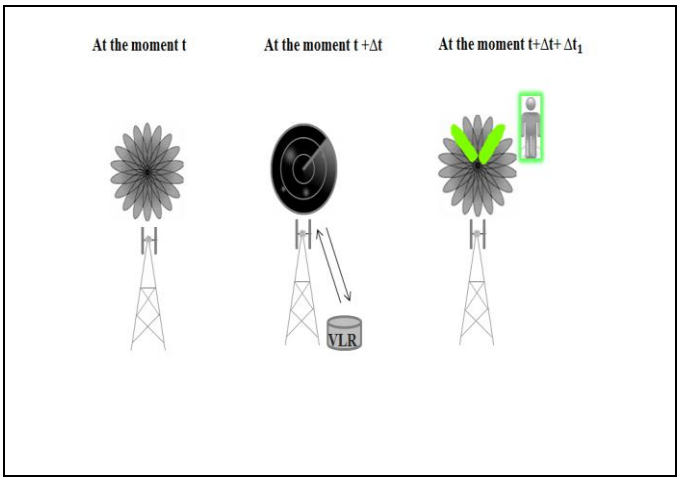


Figure 7. Starting up of the system of two dipoles with phase shifting implemented at the base station and detection of a subscriber

At the moment t+ 2Δt+ Δt_1 :

A scanning is made in all the space during Δt. This scanning allows detecting the number and the location of present subscribers in the cell. At the same time, the subscriber already detected and declared always present in the cell remains covered by the radiation.

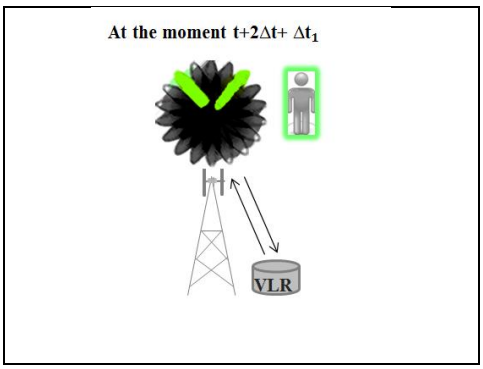


Figure 8. Covering the present subscriber and simultaneous scan for detection of new subscribers

2) Case of Two Subscribers

At the moment t+2Δt+Δt_1, and further to the end of the scan one or several other subscribers are detected present in the cell covered by BTS.
To simplify the explanation of the approach we are going to begin by handling the case of the detection of a single subscriber, and then we will see the case of several consecutive subscribers in the following paragraph. This new subscriber is covered by the radiation. The process of covering the new subscriber has to be made in duration Δt_2 whose the sum with necessary duration for the scan Δt of the cell will be in conformity with the 3GPP's standards and thus not

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

impacting on the mobile communication of the first subscriber.

- In the particular case where the new subscriber (either the new subscribers) is in the same direction, or in the symmetric direction of first subscriber; there will be no new beamforming because both beams already generated for the first subscriber cover this new subscriber as well. Thus the request of the latter subscriber is satisfied in a duration low than Δt_2.

- In case the new subscriber is in an asymmetric direction compared with the first subscriber, there will be a new beamforming. The coverage of the new subscriber consists in the generation of two lobes in the main direction and the corresponding symmetric direction via the mechanism of beamforming, realized in our case by the delay line leading to the phase shifting between the currents of both dipoles.

- Thus the request of the latter subscriber is satisfied during Δt_2 . It is important to note that both subscribers will not be covered simultaneously, but alternately, one after the other. The phase shifting and the scanning will be seamless for subscribers, because of the fact that the sum of the consecutive temporal spaces between both coverage plus duration of the scanning of the cell must be lower than 50ms limit defined by the 3GPP **[9]**.

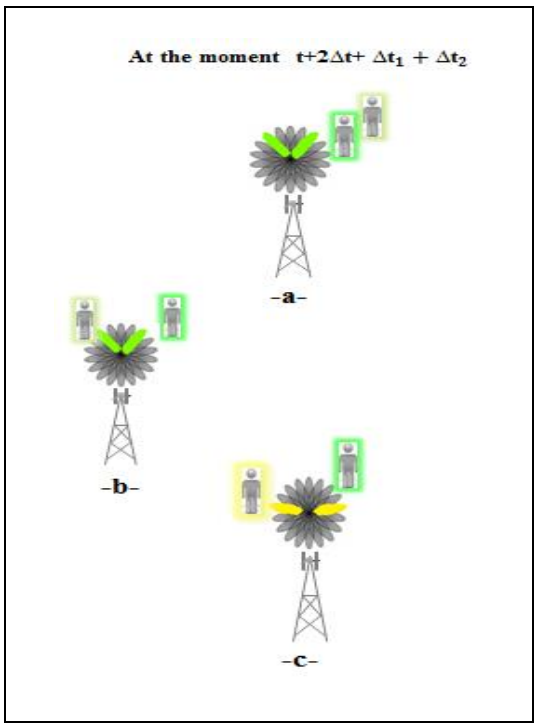


Figure 9. Covering the second subscriber in -a-similar- b-symmetric -c-asymmetric, direction compared to the first subscriber

At the moment t+ 3Δt+ + $\Delta t_1 + \Delta t_2$:

We will study the last case where the second subscriber is asymmetric with the first one.
A scanning is made in all the space during Δt.If we assum that we have only these two subscribers in the cell. After the end of this scanning, the first subscriber is once again covered by the radiation , see figure below.

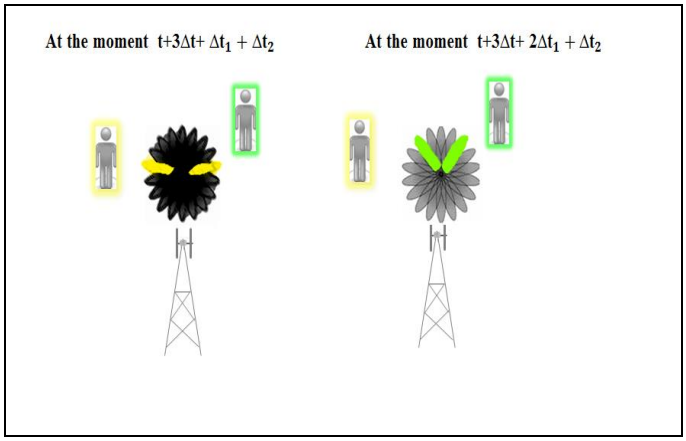


Figure 10. Covering of the first subscriber again after scanning

3) Case of Three Subscribers and More

At the moment t+ 3Δt+ + $\Delta t_1 + \Delta t_2 + \Delta t_3$:

Further to the end of the scanning in t+3Δt + $\Delta t_1 + \Delta t_2$, several subscribers are simultaneously detected quite asymmetrical to the first subscriber, we will have new beamforming. We assume that we detected three other subscribers. We will consider five present subscribers, each one of them is present in one of the five angles.
The sum of these five angles allow covering 180°. Taking into account the symmetry of radiation these five angles allow covering all the space of the cell.

- The principle is to use the table of phase shifting containing the values (φ_i), i ∈ [1, n], with n=5 total number of subscribers detected in the cell.
- The covering of five subscribers consists in the generation of two lobes in the main direction and the symmetric direction corresponding to each of five subscribers.
- The generation of both lobes is realized by the delay line leading the phase difference between the currents of both dipoles.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The values of the table of phase shifting $[\varphi_1,\varphi_2,\varphi_3,\varphi_4,\varphi_5]$ are going to be injected in the delay line with a temporal spacing between every two values of phase equal to $\Delta t_i + \Delta t$ such us:

$$\sum_{i=1}^{5} \Delta t_i + 5\Delta t \ll 50ms \quad \textbf{(19)}$$

Where Δt is the duration of the cell scanning and Δt_i the necessary time to perform the beamforming for the concerned subscriber.
Throughout the process that we defined, at the end of every Δt, a scanning will be performed. The five subscribers will be covered one after the other. Phase shifting and the scanning will be seamless, because the sum of the consecutive temporal spaces for beamforming plus duration of the scanning must be lower to 50ms limit defined by the 3GPP [9]. After the beamforming for any subscriber, he is covered by the radiation of the base station of the cell during Δt, and a scanning is launched simultaneously at the same duration.

Generally, to avoid degradation during the mobile communications during the necessary processing for beamforming as well as for the scanning, we have to be in conformity with the condition below:

$$\sum_{i=1}^{n} \Delta t_i + n\Delta t \ll 50ms \quad \textbf{(20)}$$

Where Δt is the duration of the cell scanning and Δt_i the necessary time to perform the beamforming for the concerned subscriber. N is the number of angles whose sum covers 180°.
The five chosen directions allow covering the entire cell. If after the scanning one or several additional subscribers are detected, they will necessarily be distributed on the five angles fixed at the beginning and studied, thus there will be no supplementary time delay which can influence the quality of the mobile communication as long as we are in conformity with the condition (19).
The presented example allows having a global visibility of the functioning of the system of two dipoles with phase shifting implemented in base station. In this study we grouped all the angles whose sum allows covering all the space.
If the subscribers are detected simultaneously or in succession, it will not at all impact the process of covering subscriber as long as the chosen technology (delay line, microcontroller, and system of access to the VLR) which define Δt and Δt_i will be in conformity with the condition (20).
In fact, if we assume that five subscribers are simultaneously detected, they will undergo the same processing as if they are detected successively, in this case a random sequencing will be added.

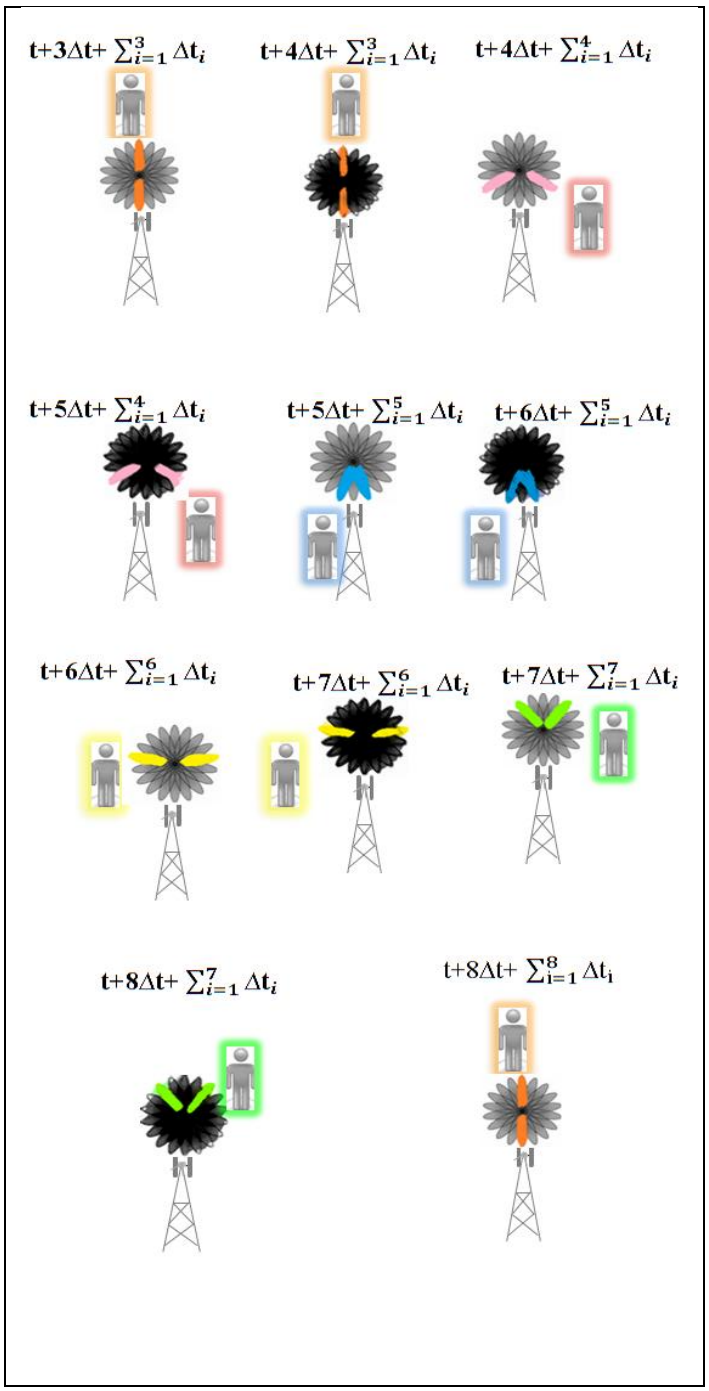


Figure 11. Chronology of covering five subscribers by a base station with two dipoles with phase shifting

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

4) Shift Towards the Simple Mode of Radiation

The adjustment of the value n will be made in the experimental limits which are Δt, the minimal time to reach the VLR, and Δt_i the duration to perform the beamforming supplied by the chosen microcontroller.

In the practice, and concretely if we consider the time perceivable for a simple subscriber, the five subscribers can all communicate simultaneously.

However, the technology being chosen well and fixed, we know well the values Δt and $\Delta t_{i,}$ which are fixed. The inequation **(20)** means finding the maximal limit of the number of the subscribers n.

$$n \ll \frac{50ms}{\Delta t_i + \Delta t} \quad \textbf{(21)}$$

Beyond the value n defined by the inequation above, the base station shift towards the simple mode. In this mode we can switch off one dipole and let a single dipole shining in the entire cell when the load of the traffic is very high.

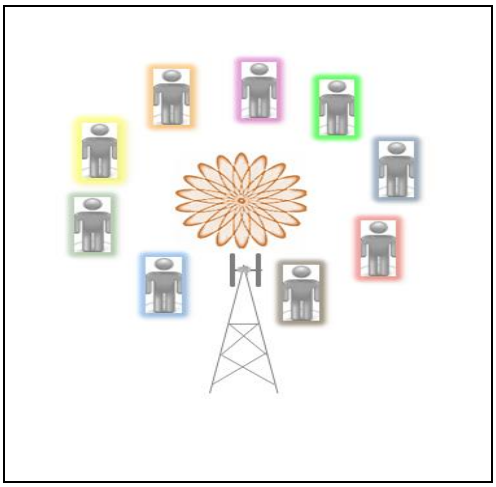


Figure 12. Simple radiation with only one dipole

IV. CONCLUSION

In this article we exposed the functioning of base station using two dipoles with phase shifting for various possible configurations concerning subscriber's distribution on the cell.

This proposed architecture allows channeling the radiation of the antennas of BTS, to reduce the exposure to not ionizing radiation; it allows optimizing the energy consumed by BTS also, thus we can reduce the costs of deployment for the operators.

We proposed a new combination of the existing smart system antenna to propose at the same time simple and effective model to strengthen the application of the smart antennas in base stations.

This system joins the database of the green technologies contributing to reduce the excessive radiation of the antennas which can be harmful to the subscribers' health in particular and the other people present in the coverage area generally.

The main profit of such a system is the increase of the user capacity of the cellular system. Indeed, the possibility of directing the beam of an antenna, without any mechanism of rotation, added to the possibility of obtaining beams having an important gain and a narrow opening, allows to make a vast coverage of the area and to follow the movements of a user inside the same cell by minimizing the noise and the interferences.

As a perspective of the system proposed. We can implement this system for a femtocell. It is possible to apply this solution in the 3G for data as well.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] *ICNIRP "Guide pour l'établissement de limites d'exposition aux champs électriques, magnétiques et électromagnétiques-Champs alternatifs (de fréquence variable dans le temps, jusqu'à 300 GHz) "Cahiers de notes documentaires - Hygiène et sécurité du travail N° 182 - 1er trimestre 2001*

[2] *Rapport no 52 de l'OPECST, page 38*

[3] *ANRT- "Rapport sur l'appréciation des conditions d'exposition aux rayonnements des champs électromagnétiques non ionisants" - Version I Février 2002*

[4] *Royaume du Maroc-Ministère de santé "Valaurs limites d'exposition aux champs électromagnétiques émis par les équipements utilisés dans les réseaux de telecommunication ou par les installation radioélectriques" - Circulaire N°21- -22 mai 2003*

[5] *J. Butler and R. Lowe, "Beam-Forming Matrix Simplifies Design of Electronically Scanned Antennas", Electronic Design, pp. 170-173, April 12, 1961.*

[6] *JUMIRA Oswald, ZEADALLY Sherali "Efficacité énergétique des réseaux sans fil "édition Lavoisier, 2013, ISBN : 2746295334, 9782746295339*

[7] *Constantine A. Balanis "Antenna theory analysis and design «third edition a John Wiley & Sons, inc., publication p283 -333*

[8] *G. Faillon and G. Fleury. Tubes pour hyperfréquences.Techniques de l'ingénieur, traité Electronique, Référence E 2 355.*

[9] *Borko Furht, Syed A. Ahson "Long Term Evolution: 3GPP LTE Radio and Cellular Technology"*

An Efficient Model to Automatically Find Index in Databases

Mohammad H. Nadimi-Shahraki
Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran

Rezvan Shahriari
Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran

Mohamad Davar Panah Jazi
Department of Computer Engineering, Foulad Institute of Technology, Fouladshahr, Iran

Abstract— The high volume and increasing complexity of data in large organizations requires optimization methods for fast access to data in a database. Choosing the most appropriate index for the database is essential to optimization; the growing need of more complicated indexing makes reliance on current database administrators insufficient. Employing a skilled database administrator to work with complex database management systems is expensive and not an option in many organizations. A more cost-effective approach is use an automated data management system to find the index. Several such approaches have been developed to date. Although there have been many methods based on data mining for finding a proper index, their run time is still unacceptable. Particularly, in very large databases, faster methods for finding proper indexes are needed. In this paper, an efficient model is proposed to automatically find index in databases using maximal frequent patterns. The proposed model is evaluated by conducted benchmark dataset TPC-H to compare with previous methods. The evaluation results show that using the proposed model can decrease the time of finding indexes in databases.

Keywords- Optimization, automatic indexing, maximal frequent patterns

I. INTRODUCTION *(HEADING 1)*

Techniques for fast access to data in a database are required to optimize a database. In most professional environments, relational databases constitute the core of programs and business services. As data volume and the number of users and applications based on relational databases increase, organizations must ensure the optimal performance and response of their own databases. A query processing engine in a relational database enables users to implement maximal performance and optimize their programs. These processing engines require optimal performance based on an index to quickly locate data without having to search every row in a database table each time. Indexes can use one or more columns of a database table to find data more quickly [1]. With the increased complexity of indexes and the decreasing adequacy of a database administrator to find data using the index, the alternative of automatically finding indexes in the database has been suggested. Heuristic methods and expert systems have been used to automatically find an index [2, 3]. The Auto Admin project developed a method that finds indexes in Microsoft SQL Server using a physical design that is completely automated [1]. Data mining techniques have been increasing tested to automatically find an index [4-7], but may not be able to produce one in less time. The present study developed an effective way to decrease the time for automatic index selection using maximal frequent patterns.

Section 2 describes the problem and previous attempts to find indexes automatically. Section 3 describes the proposed method. The experimental results are discussed in Section 4. Section 5 summarizes the study.

II. *PROBLEM STATEMENT AND LITERATURE REVIEW*

A number of studies have focused on finding the index in a database automatically. The Auto Admin project developed a new technique that enables the system to select an index automatically in Microsoft SQL Server in a physical design. SQL Server initially receives a set named "workload" composed of multiple queries with the overall objective of suggesting an optimal set of indexes. This method has been implemented only on SQL Server database management system.

Prior to Auto Admin, three approaches were used to solve this problem. The first is the textbook solution in which semantic information such as resource constraints and basic statistics are used [2]. In this solution, the creation and design of such a scheme in the database was insufficient, because the value of workload information was overlooked.

In the second approach, the expert system used a means of coding rules. This approach connected with the workload, but

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

failed to create a connection with the query optimizer. The system failed for two reasons. First, the optimizer required appropriate index selection for use by the optimizer. The second is that this tool applied an index used in the query optimizer, where incompatibility between the tool and the optimizer can be problematic when choosing the optimal index.

The third approach is an index selection tool for estimating optimizer costs to determine the efficiency and sufficiency of Auto Admin project indexing. This system prevents synchronization between the tool and the query optimizer [8].

Gradual advances in data mining allow the use of frequent patterns to find indexes automatically. Frequent patterns are subsets that repeat in a data set with a frequency greater than or equal to that specified by the user. For example, a set of items (like milk and bread) usually found in the base transactions of a supermarket forms a set of frequent items [9]. Han (2000) developed a method called FP-Growth [10] by which the complete set of frequent items can be accessed without creating candidates. This method is based on divide and conquer strategies where, in a single scan, all items available in the transaction base are counted and arranged according to the transaction frequency in descending order. This list of frequent items can be compressed into the form of a frequent pattern tree (FP-Tree) containing all information in the database and used for look-ups. The FP-Growth algorithm converts large frequent patterns into smaller frequent ones recursively and then attaches suffixes to the specific patterns. One major challenge in this process is that the algorithms provide too many frequent patterns as results.

When a small minimum support is chosen, it becomes more evident because all subsets of a frequent item are also frequent. A large frequent item in the database may result in the exponential growth of the number of frequent items (a subset of the set of large items) in the transaction database. Maximal frequent patterns have been proposed to address this problem, where P is the pattern in a data set and D is the maximal frequent pattern, if P is frequent and there is no frequent super pattern for P [11]. The main problem to accessing a frequent pattern is whether it is a maximal pattern or not.

One main solution to this problem is to keep the TID of the patterns and index by hashing TID values of patterns. This method is used in the CHARM algorithm [12]. Another is to store the accessed patterns in a pattern tree (similar to FP-Tree). CLOSET+, AFOPT, FPClose and FPMax techniques use this approach in their algorithms [13-16]. Maximal frequent patterns are viable alternatives to looking up frequent patterns because the analytical power of all frequent patterns is identical, but the end result is far lower than the number of patterns in all frequent patterns. This means that looking up maximal frequent patterns is highly scalable and extensible.

III. PROPOSED SOLUTION

Proposed method for automatic selection of indexes in databases considers whether it is possible to decrease the time cost. One effective method that accurately finds indexes in databases uses frequent patterns (algorithms, FP-Growth) [5]. The proposed method for decreasing time cost is to select multiple-column and single-column indexes in the database using maximal frequent pattern algorithms. This means that the maximal frequent patterns are found among transactions produced by analyzed queries and are used to compose the indexes. Figure 1 summarizes the proposed model stages.

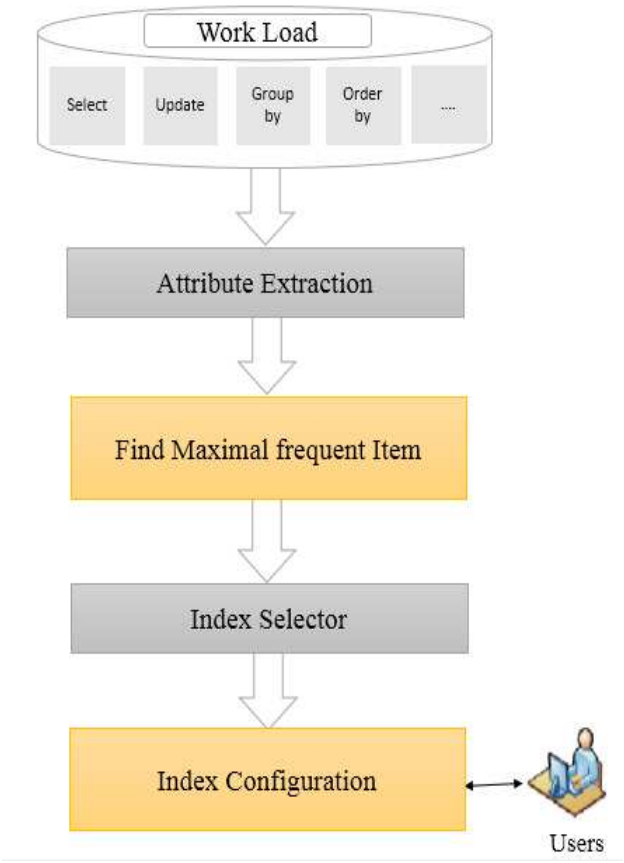


Figure 1: The proposed model.

Assume a set of available queries in a database as shown in figure 2. The columns to be indexed meet one of these conditions or is used in Order-by, Group-by or collective functions. An update requirement can also be indexed. Given these index selection criteria, the workload and transactions can be obtained from the queries. Table 1 lists the relevant transactions from a built matrix

The proposed method finds the indexes automatically and the patterns obtained generate index able columns. The maximal frequent patterns obtained from a transactional table are: {B}, {C, P}, and {F, C, A, M}. {B} is a single index and{C, P} and {F, C, A, M} are multiple indexes. This method decreases the time required for selecting an index for the database.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
Q1: select * from T1, T2 where F between 1 and 10 and C=A and M>100 order by P
Q2: select * from T1, T2 where A like '%50%' and B=5 and C<100 and F=M
Q3: select * from T1 where B>2 and F in (3, 2, 5)
Q4: select * from T1, T2 where B>3 group by C having sum (P)>2
Q5: select * from T1, T2 where A=30 and F>3 group by C having sum (P)>2 and M=10
```

Figure 2: Queries available in a database.

The proposed method finds the indexes automatically and the patterns obtained generate index able columns. The maximal frequent patterns obtained from a transactional table are: {B}, {C, P}, and {F, C, A, M}. {B} is a single index and{C, P} and {F, C, A, M} are multiple indexes. This method decreases the time required for selecting an index for the database

Table 1: Transactional database obtained from queries.

Items	Transaction
f, c, a, m, p	T_1
a, b, c, f, m	T_2
b, f	T_3
b, c, p	T_4
a, f, c, p, m	T_5

IV. EVALUATION

In this section the proposed method is experimentally evaluated conducted by standard benchmark TPC-H which is used by most existing literature comparison for automatic selection of the indexes. All trials in the present study were implemented on Windows 7 using the CPU Core i7 2.2 GH with 8 GB memory. The experimental results shown in Figure 3 indicate that the automatic finding and creation of indexes using the proposed method required less time than previous methods and made use of maximal frequent patterns.

V. CONCLUSIONS

This study proposes a new method to automatically find indexes in a database as an advanced query optimizer for database management systems in which multiple and single indexes are selected using maximal frequent pattern algorithms. Frequency tests were conducted on the TPC-H benchmark to test the results of the method. Results show that the proposed method decreases the time required for finding indexes over that of previous methods.

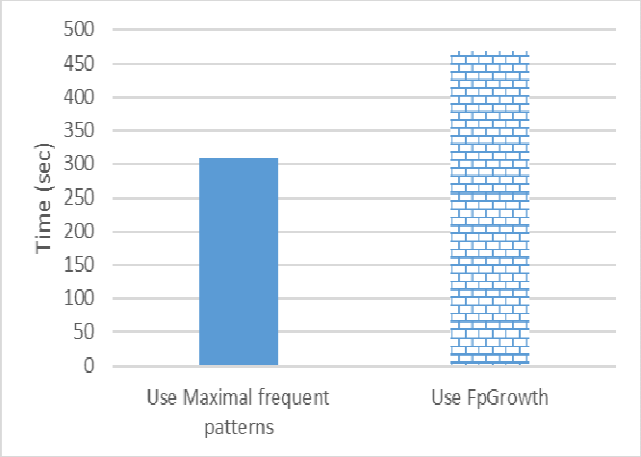


Figure 3: Time required finding and creating indices.

VI. REFRENCES:

[1] S. Chaudhuri and V. R. Narasayya, "An efficient, cost-driven index selection tool for Microsoft SQL server," in VLDB, 1997, pp. 146-155.

[2] W. N. Venables, B. D. Ripley, and W. Venables, Modern applied statistics with S-PLUS vol. 250: Springer-verlag New York, 1994.

[3] P. Corrigan and M. Gurry, "Oracle performance tuning-database management systems: covers versions 6 and 7," ed: O'Reilly, 1993, pp. I-XXXV,1-603.

[4] R. Agrawal and G. Psaila, "Active Data Mining," in KDD, 1995, pp. 3-8.

[5] K. Nimkanjana, S. Vanichayobon, and W. Wettayaprasit, "Auto-Indexing Selection Technique in Databases under Space Usage Constraint Using FP-Growth and Dynamic Programming," in Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on, 2008, pp. 932-935.

[6] M. Zaman, J. Surabattula, and L. Gruenwald, "An auto-indexing technique for databases based on clustering," in Database and Expert Systems Applications, 2004. Proceedings. 15th International Workshop on, 2004, pp. 776-780.

[7] K. Aouiche, J. Darmont, and L. Gruenwald, "Frequent itemsets mining for database auto-administration," in Database Engineering and Applications Symposium,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2003. Proceedings. Seventh International, 2003, pp. 98-103.

[8] S. Chaudhuri and V. Narasayya, "Self-tuning database systems: a decade of progress," in Proceedings of the 33rd international conference on Very large data bases, 2007, pp. 3-14.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, 1993, pp. 207-216.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD Record, 2000, pp. 1-12.

[11] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in ACM Sigmod Record, 1998, pp. 85-93.

[12] M. J. Zaki and C.-J. Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining," in SDM, 2002, pp. 457-473.

[13] B. Goethals and M. J. Zaki, "Advances in frequent itemset mining implementations: report on FIMI'03," ACM SIGKDD Explorations Newsletter, vol. 6, pp. 109-117, 2004.

[14] G. Liu, H. Lu, J. X. Yu, W. Wang, and X. Xiao, "AFOPT: An Efficient Implementation of Pattern Growth Approach," in FIMI, 2003.

[15] J. Wang, J. Han, and J. Pei, "Closet+: Searching for the best strategies for mining frequent closed itemsets," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 236-245.

[16] G. Grahne and J. Zhu, "High performance mining of maximal frequent itemsets," in 6th International Workshop on High Performance Data Mining, 2003.

.

An Overview and Comparison of Hierarchical P2P-SIP Networks

Abel Diatta [1*], Ibrahima Niang [1], Mandicou Ba [2]

[1] *Département de Mathématiques et Informatique*
Laboratoire d'Informatique de Dakar (LID)
Université Cheikh Anta Diop de Dakar
Dakar, Sénégal
{abel.diatta,ibrahima1.niang}@ucad.edu.sn
[*] Corresponding author: Abel Diatta

[2] *Université de Reims Champagne-Ardenne*

Laboratoire CReSTIC EA 3804 - Équipe SysCom
Reims, France
mandicou.ba@univ-reims.fr

***Abstract*—P2P-SIP networks are emerging distributed communication technology introducing Scalability and Robustness. These networks propose to use a peer-to-peer network for user registration and user location in Session Initiation Protocol (SIP)-based voiceover- IP (VoIP) networks.**

Nevertheless, P2P-SIP systems are based on a pure flat DHT design which provides a uniform distribution of (equal) peers and resources that assures scalability and load balancing. Recent research works have proposed to organize P2P overlays into hierarchical architecture called HP2P-SIP. These later are motivated by supporting sophisticated search requirements, separating categories of use (content, personal communications,), scaling, etc. However, these works did not resolve the problem of overloading the bandwidth, which is still a head-case.

In this paper, after highlighting the many challenges of hierarchical architectures, we show how to avoid overloading the physical network by acting on the overlay network. To do this, we have a physical network of size $\mathcal{N}$ with a number of super nodes ($\mathcal{N}_{SN}$). Our solution shows how many levels k we must build our overlay network so that in the physical network, the number of messages generated by the lookup nodes does not exceed a value x, initially fixed.

***Index Terms*—HP2P-SIP, Costing, Iterative, Recursive, Semi-recursive, Bandwidth overload**

I. Introduction

In recent years, Session Initiation Protocol (SIP) has emerged as a signalling protocol in multimedia communication networks [1]. This protocol is becoming as a promising signalling protocol for the next generation network like IP Multimedia Subsystem (IMS). However, current SIP systems implementations are typically based on the client-server architecture. This exposes some problems and disadvantages including high input costs, poor scalability, and disaster recovery capacity. Moreover, traditional centralized server based SIP system is vulnerable to several problems like performance bottleneck and single point of failure.

Considering their good robustness, and high extensibility, Peer to Peer (P2P) systems are expected to be a perfect method to solve the bottleneck in client/server pattern. P2P networks are favored for their adaptation, self-organization, and decentralized control. Therefore, Peer-to-Peer Session Initiation Protocol (P2P-SIP) is proposed to leverage the inherent advantages of Distributed Hash Table (DHT) such as low maintenance, scalability, robustness, and fault tolerance. The main reason is that they are no centralized server, and its self-organizes nature [2], [3], [4].

Nevertheless, most of the existing P2P-SIP systems are based on a pure flat DHT design which provides a uniform distribution of (equal) peers and resources that assures scalability and load balancing in the overlay network. In this case, all participating peers are considered equal in functionality. However, nodes do not have the same capabilities. In other words, some of them have more processing power and storage than others. Therefore, by putting them together, low capacity nodes may slow down the operation of high capacity nodes.

To overcome this problem, recent research works have proposed to organize P2P overlays into hierarchical architecture [3], [5], [6], [7], [8], [9]. In this new hierarchical architecture, nodes are classified into super peers and ordinary peers [7], [9], [6]. Super peers provide server functions registration and location to ordinary remote peers [1]. However, despite the hierarchical organization, many problems remain unresolved like the network overload, overload of super nodes, crossing NAT, ... To alleviate these problems, it is important to carefully choose the routing method. In addition, as some mechanisms are more expensive than others in terms of the number of generated messages, as a result, the time to locate may be. However, the latter is a critical criterion for real time communication services. Many works [10], [11], tried to improve these metrics (network overload, overload of super nodes, ...) without being able to limit them. In [11], autors do a comparison between flat and hierarchical system desing in termes of costs. While in [10], Baumgart et *al.* treated these metrics like the consumption of bandwidth, routing mechanisms in different protocols (Chord, Bamboo, Pastry, Koorde, ...).

In [10] they simply showed which of these protocols was the best, but failed to put in place a strategy to limit occupancy of bandwidth. Many existing works [3], [8], [12]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

have improved the overload of the bandwidth by offering hierarchical architectures instead of flat architectures. But, once on hierarchical architectures, they do not have strategy to make fluid bandwidth.

The special feature of our solution (in this paper) is that we show how to build an overlay network by limiting the number of messages that can pass through the physical network, and that, whatever the size of the physical network. To do this, we have a physical network of size $\mathcal{N}$ with a number of super nodes ($\mathcal{N}_{SN}$). Our solution shows how many levels k we must build our overlay network so that in the physical network, the number of messages generated by the lookup nodes does not exceed a value x, initially fixed.

The rest of the paper is organized as follows. P2P-SIP concepts and Chord-based overlay designs are presented in section II. In section III, we propose a classification of hierarchical P2P-SIP overlays architectures. A comparison between these overlays is also proposed. In Section IV, we calculate the exact cost in terms of messages generated in different hierarchical architectures. Finally, in Section V, we present a conclusion and open issues.

II. P2P-SIP Design

In this section, we take a brief overview of P2P and P2P-SIP systems, but also the most popular and most widely used protocol [8], [13] namely Chord.

A. P2P Systems

P2P networks are distributed systems in nature, without any hierarchical organization or centralized control [14], [4], [13]. They have many advantages, such as scalability, robustness, fault tolerance, etc. Every communication entity in these networks plays an equally important role, which is named peer or node. Node has both client and server features. Therefore, a node in network is not only a client, but also provides functionality as servers to respond to other clients [15]. They are mainly organized in structured and unstructured manner [4].

On the one hand, unstructured networks rely on flooding techniques [16]. In this system, all nodes have the same capability to serve other clients. In addition, the network does not have any logical structure. There is one distinct character of this system: while node wants to search other resource or obtain any service, it will flood the request to all of its neighbour nodes. Flooding-based systems don't scale well because of the bandwidth and processing requirements they place on the network, and they provide no guarantees as to lookup times or content accessibility.

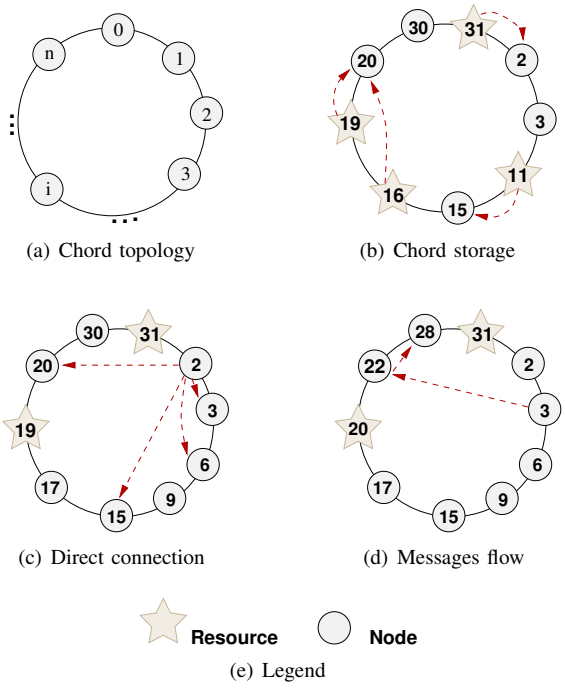
On the other hand, structured networks impose techniques to tightly control the data placement and topology within the network, and currently only support search by identified. In many scenarios, the increased search efficiency makes structured networks preferable to the widely deployed unstructured networks. There are several flavors of DHT [13] (Chord [4], Pastry [15], Kademlia [10], Bamboo [16], Tapestry [5], etc) each with some advantages over another.

B. Chord-based P2P-SIP overlay design

We start this section by defining overlay network. To the best of our knowledge, one of the best definition of overlay network is given by Xianghan et *al.* in [2]: *"An overlay network is virtual network of nodes and logical links that is built on top of an existing network with the purpose to implement a network service that is not available in the existing network"*.

In P2P systems, a DHT table has been introduced in order to provide efficient location and retrieving operations. A DHT is a decentralized and distributed system where all nodes and resources are identified by unique keys. Several DHT algorithms have been proposed in the literature like Chord [4], Bamboo [16], Pastry [15], Tapestry [5] etc. However, in the remainder of this paper, we focus our work exclusively on the most popular DHT : Chord algorithm [4]. Therefore, in the following subsection, we specify Chord-based P2P-SIP overlay. In fact, this latter has been suggested as a mandatory overlay to support P2P-SIP communication.

As specified by Furness et *al.* in [17] and in Xianghan et *al.* in [2], in Chord overlay, peers and resources are structured into a ring, as illustrated in Fig. 1(a). Peers and resources are represented by integers $\mathcal{N}odeID/\mathcal{R}esourceID$. Each peer stores a pair of values (id, $value$), where id represents the peer or resource $\mathcal{ID}$, and $value$ represents the peer address information or the data storage. Peer and resource $\mathcal{ID}s$ can be obtained by consistent hashing using the cryptographic *SHA-1* algorithm [18], [8], [13]. The peer $\mathcal{ID}$ is produced by hashing the IP address of the particular peer, and the resource $\mathcal{ID}$ is obtained by hashing the data value. The $\mathcal{R}esourceID$ is stored in the first peer which $\mathcal{ID} \geq \mathcal{R}esourceID$ (Fig. 1(b)).



(a) Chord topology (b) Chord storage

(c) Direct connection (d) Messages flow

(e) Legend

Fig. 1: Lookup and Storage in Chord architecture

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

In addition, each peer has a Finger table, in order to ensure the routing information records. As show by Furness et *al.* in [17], the Finger table records $\log_2(\mathcal{N})$ successors, where $\mathcal{N} = 2^m$ is the number of peers in the overlay and m represents the number of addressing bits (Fig. 1(c)). The i^{th} successor $\mathcal{ID}$ of a peer which $\mathcal{ID}$ is $\mathcal{P}$ is determined as follows :

$$SuccID(i) = (\mathcal{P}+2^{i-1}) \mod 2^m, with\ (0 < i \leq m). \quad (1)$$

Each peer contacts periodically its successors for updating its own Finger table. It also contacts the predecessor that is the previous peer in the identifier circle. This is useful when a peer leaves the ring and asks the previous peer to update its Finger table.

During the transfer information, Chord routes the data by sending messages to the next successor nearest to the destination identifier. As illustrated in the example of Fig. 1(e), peer 3 attempts to locate peer 28. Firstly, it checks its finger table records, chooses a successor (peer 22) nearest to the destination, and then send a request to this successor. Similarly, peer 22 checks its own finger table and forwards the message to its successor (peer 28).

C. P2P-SIP systems

Recent research activities have proposed to use a peer-to-peer network for user registration and user location in Session Initiation Protocol (SIP)-based voice-over-IP (VoIP) networks [4], [3], [19]. The main motivation for P2P-SIP is their higher robustness as well as their easy configuration and maintenance (compared to client-server SIP).

1) P2P-SIP requirement statements: In P2P-SIP architectures, for efficient operation, it is necessary to satisfy some requirements that have been emerged separately by several research [2], [4], [5], [6], [15], [16], [18], [20], [21], [22], [23], [24], [25]. In this section, we summarize all of these requirements. Note that many of them were reported by Zheng et *al.* in [2] and Singh et *al.* [21]. In our case, a broad overview (in section 3) in the existing literature has allowed us to see what are the requirements for which, solutions have been found and those that are not yet resolved by the community of researchers. The goal of this, is to help researchers to move towards the outstanding issues.

- ***R1: Availability, Efficiency, and Stability***: P2P-SIP system should allow to recover a resource, if it exists. The system must also be able to return to a stable state after the arrivals and departures of nodes [15], [22], [24].
- ***R2: DHT overlay flexibility***: A DHT should be easy to implement. It must also be able to take into account other existing DHT (as Chord, Pastry, Bambo, Kademlia, ...) but it also be applicable in the future [6], [16].
- ***R3: Inter-working***: Good DHT must allow interconnection with other systems such as PSTN networks, but also with the next generation networks as IP Multimedia SubSystem (IMS) [5], [20].
- ***R4: NAT and firewall Traversal***: The establishment of a DHT must consider the fact that some users are behind NAT and firewalls. It is important to find strategies to cross them. Some protocols exist for this purpose (STUN, TURN, ICE) [22], [25].
- ***R5: P2P-SIP Client Protocol***: According to [2], the designed peer protocol should contain the client protocol to support the legacy devices that participate the P2P-SIP overlay but do not make contributions due to a lack of the support in DHT algorithm or limitation of devices capability (e.g energy, CPU processing power, bandwidth,...).
- ***R6 :Security requirement***: Security is one of the biggest challenges for P2P and P2P-SIP systems. The system must have data protection mechanisms that users share [15], [20], [22].
- ***R7:Fault tolerance***: Since node failures can happen at any time, the system must provide good strategies to restart the system after a node failure. Fault-tolerant solutions must consider physical and timing faults of nodes [23].

2) P2P-SIP Challenges: As for the requirements, to establish a good P2P-SIP system, it is important to address some challenges.

- ***C1: Resource Lookup Delay***: In an instant communication environment, the time to locate a node is an important factor. Trying to reduce this delay is significant challenge [4], [5], [21].
- ***C2: Network Address Translation (NAT)***: Most of the P2P nodes may be behind a NAT or Firewall. There must be some relay in between them with a public IP address in order to establish end to end communication . This is one of the most important challenges for P2P-SIP network [25], [22].
- ***C3: Node Heterogeneity***: As the nodes do not have the same capabilities (bandwidth, CPU, storage, ...), the system must accept heterogeneous nodes [6], [21], [24].
- ***C4: Security Issues and Trustworthiness of Peers***: Security in distributed P2P communication system is another of the major challenges. Security issues concern user identification, authentication and trustworthiness [6], [15].
- ***C5: Fault Tolerance***: When a node fails, the system must quickly return to normal [23].
- ***C6: fluidity of bandwidth***: To make effective for example the searches for resources, it is essential to have fluid networks in which the data can circulate without too much waste of time [21].
- ***C7: No clutter of super nodes***: Also for a better circulation of the information and a good data processing, the super nodes must not be too much blocked [4], [5], [16].

III. Hierarchical P2P-SIP Overlay Architectures

In this section, we do an width overview on the existing hierarchical P2P-SIP architectures and then we compare them. In P2P-SIP architectures, all nodes, whether large or small capacities are on the same level. This can cause some delay in processing requests. Because in the reality, all nodes don't have the same processing power. At the opposed, in the case of hierarchical nodes, only those with large capacities (called

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

super nodes) act as registration servers, redirect servers and location servers [24], [23], [12]. The other nodes with low capacities are called ordinary nodes. Therefore, this allows better speed up the processing of requests and save time. Different types of hierarchical architectures exist in the literature:

- the architectures on which super nodes are in the same level and the ordinary nodes are attached to them [8], [7]
- the architeures on which super nodes are in the same level and ordinary nodes are organized into sub-levels which are connected to super nodes [3], [5], [6]

In the first case we have an overlay with a single domain. In the second case we have an overlay network multi-domain. In this second case it is possible to use a DHT per domain as for example in [22] and in [12]. This has a positive effect on the speed of research resources, but also on the use of bandwidth.

A. HP2P-SIP and overlays organization

In this sub-section, we go through the different papers on hierarchical architectures to highlight strengths and weaknesses of different architectures, namely architectures single domain and multi-domain architectures. So we can identify the challenges ahead.

1) Single Domain Organization: HP2P-SIP single domain includes two different overlay architectures: one layered overlay and the multi-layered overlay [7], [24].

a) One-layered overlay: One layered overlay is the default architecture for HP2P-SIP based on the concept of Super Nodes (SN) and Ordinary Nodes (ON). Only the Super Nodes form the overlay and Ordinary Nodes are connected to them.

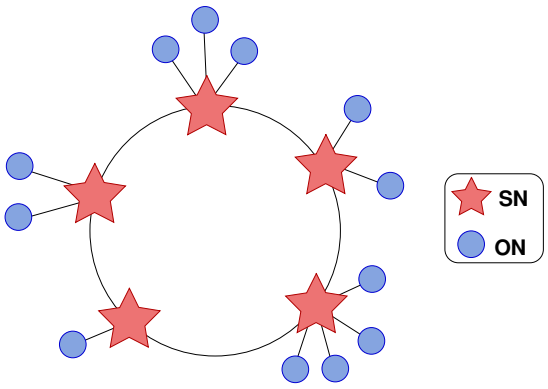


Fig. 2: One layered overlay architecture

Singh et *al.* in [18] propose one overlay architecture based on the concept of ordinary node and super node with Chord as the underlying DHT. Their architecture supports basic user registration and call setup. Their paper fails in practical challenges because they just give an overview such as call setup latency, NAT traversal security issues.

In [22], autors define one hierarchical architecture termed manageable approach. In their design, only SIP servers can be super nodes and form the overlay. The others nodes called SIP clients can be standard traditional SIP clients or P2P-SIP clients. For some practical challenges they propose a solution for NAT traversal, Call set up delay. Their paper fails to address security problem and reliability of SIP signaling and fault tolerance issues.

b) Multi-layered overlay: They consist of several layers. The uppermost layer is formed by super nodes and the other nodes comprise other layers depending on their capabilities [24]. Smaller capacity nodes occupy the lowest layer. In multi-layered overlays we can have two or more than two layered overlays. Nodes that have low capacity form the low level overlay, and for those have more (most powerful and stable) build the top level overlay . Routing and maintaining different overlays is done by a proper DHT in each layered, like CAN, Kademlia, Chord...

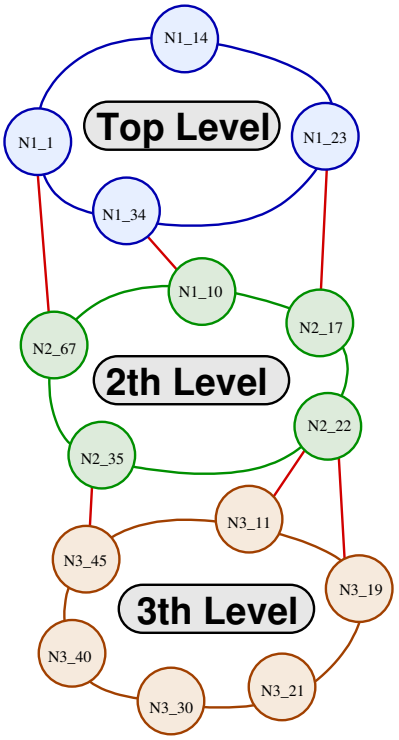


Fig. 3: Tree-level hierarchical overlays

Diané et *al.* in [23] propose an architecture for an efficient Fault Tolerant approach based on three levels. The first level is formed by nodes termed Light supernodes, the second level is built by super nodes and the low level includes the ordinary nodes. Their architecture allows the accessibility of ordinary nodes during the physical failure and the timing failures of super nodes and to reduce the localization delay of Remote Ordinary Nodes (RON), when its attached super node is breakdown. But their paper fails to propose a mechanism for communication between Light super nodes and super nodes. In addition, they do not show how a super-node knows his attached light super-node. They do not consider the security, NAT traversal, interworking....

In [24], Le et *al.* suggest a hierarchical and breathing P2P-SIP architecture composed of more than two suboverlays with different physical capabilities and system availability.Their proposal succeeded to reduce the call setup delay by different

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

types of lookups. They do not give solution as the following issues security, NAT traversal, Interoperability with others system, Fault Tolerance for quality of services.

2) Multiple Domains Organization: HP2P-SIP architectures multi-domains are those constituted by several domains connected with a main domain. According to the DHT algorithm used, they can be classified into two categories: One used DHT schema and Multiple used DHT schema.

a) One-used DHT schema: In this case, one DHT algorithm is used in order to assure maintenance, routing, locating....

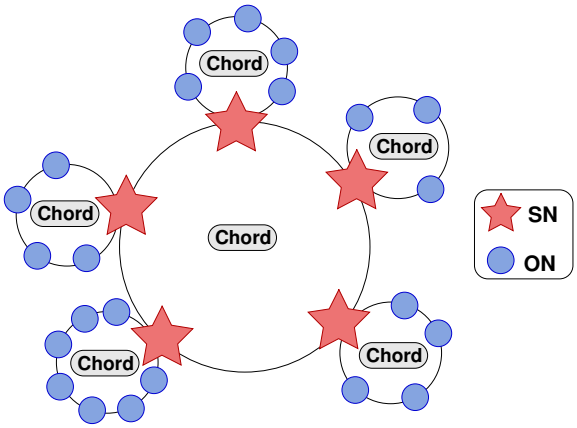


Fig. 4: Hierarchical P2P-SIP architecture Multi Domains

Ma et *al.*, in [4], suggest an architecture named DChord where they divided the overlay into Sub-nets and Main-net. Only super nodes form Main-net while ordinary nodes built the Sub-nets according to their IP address. Each super node of a Sub-net belongs to the Main-net. In their paper they used Chord as the only routing algorithm in Main-net and Sub-nets. They succeeded to improve the accuracy, efficiency and NAT traversal. Therefore their paper fails to overcome problems related to security and fault tolerance. They also overload the super nodes because they pass by them to access to the main-net.

In [15], Huang Yongfeng et *al.* conceive a hierarchical architecture formed by groups for the security of P2P-SIP networks. Every group is managed by a super node which stoques the information of the ordinary nodes. Their approach uses a cryptographic method for the authentification of nodes. For more details, the reader can consult [15]. The advantage of this algorithm (one-way accumulator) is that it supplies a too stiff method of security by a rigorous selection of peers authorized to connect in another peer. Only peers having satisfied a certain condition are selected. However their algorithm is a big consumer of bandwidth since when a node arrives or leaves the network, they recalculate the keys of the set of nodes and redistribute them throughout the network.

b) Multiple used DHT schema: In these types of architectures the DHT used in the main-level is not necessary the same that those used in sub levels. We have good examples of these architectures in [6].

In [6], Md. Saqiful et *al.*, to overcome problems encountered in combined P2P-SIP, suggested a architecture formed with different overlays. Their architecture is built on:

- a main domain overlay with Super Nodes where Bamboo is used as underlying algorithm
- different others domain overlays with Ordinary Nodes where Chord, tapestry, CAN, Pastry are used.

To overcome problems encountered in combined P2PSIP, particulary challenges, they take in many proposals. Their proposal takes into account NAT Traversal, connectivity problem, lookup delay, node heterogeneity, security....By cons, it greatly overload the super nodes because all the sub domains pass by them to access to the main domain.

In the same perspective, an architecture HP2P-SIP using several DHTs is proposed in [5]. Juwei Shi et *al.* designed an architecture with a main domain (the Higher Level Overlay - HiLO) to which are attached sub-domains (the Lower Level Overlay - LoLO). The Hilo peers are selected according to their high capacities. The interconnection between the HiLO and LoLO is assured by HiLO Peers. These are on horseback between the HiLO and LoLO. Their architecture takes care the efficiency, the call setup latency but also the NAT traversal by using the ICE protocol. However their approach has the inconvenience to overload too much Hilo Peers because all the peers of the LoLO (LoLO Peers) pass by them to access to the HiLO. They do not consider the security requirement.

In the same way in [16], the authors set up groups of networks overlay called communities and a main overlay. Every community can have its own DHT with a responsible peer which is the creator of this community. Every community possesses a community identifier. A couple $<$ *community_identifier*, *node_identifier* $>$ is assigned to every node. Their system is compatible with any DHT algorithm, thus it is flexible. It allows to accelerate the locating nodes. On the other hand it increases the load of the communities peer creators.

B. Comparison of HP2P-SIP overlays

In this section we are going to do some comparisons (TABLE I and TABLE II) of different HP2P-SIP overlays. First we compare how different HP2P-SIP overlays are trying to make more efficient P2P-SIP for the implementation of requirements and challenges.

The purpose of these comparisons is to highlight the many challenges faced in HP2P-SIP architectures and thus help researchers to easily find the problems that must be addressed in order to provide solutions.

Legend of the two tables (TABLE I and TABLE II):
No : means that the requirement or the challenge is not taken into account
Yes: means that the requirement or the challenge is raised

1) Comparison of HP2P-SIP Overlays Single domain: This table (TABLE I) shows that certain requirements, such as the inclusion of heterogeneous nodes ($C3$), are mature in the literature. In contrast, the flexibility of DHTs ($R2$) for example, or the P2P-SIP Client ($R5$) always remain strong requirements for P2P-SIP networks. It is the same for fault tolerance ($R7$) (especially for delay faults) that is hardly taken into account in the literature. Similarly, challenges such

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I: Comparison of Single domains HP2P-SIP architectures

		Single domain			
		one-layered		*multi-layered*	
		[18]	[22]	[23]	[24]
Requirements	$R1$	No	Yes	Yes	Yes
	$R2$	No	No	No	No
	$R3$	No	Yes	No	No
	$R4$	No	Yes	No	No
	$R5$	No	No	No	No
	$R6$	No	Yes	No	No
	$R7$	No	No	Yes	No
Challenges	$C1$	No	Yes	Yes	Yes
	$C2$	No	Yes	No	No
	$C3$	Yes	Yes	Yes	Yes
	$C4$	No	No	No	No
	$C5$	No	No	Yes	No

TABLE II: Comparison of multi-domains HP2P-SIP architectures

		Multi-domains					
		One-used DHT			Multiple-used DHT		
		[4]	[25]	[15]	[6]	[5]	[16]
Requirements	$R1$	Yes	No	Yes	Yes	Yes	No
	$R2$	No	No	No	Yes	No	Yes
	$R3$	No	No	No	No	Yes	Yes
	$R4$	Yes	Yes	No	No	Yes	No
	$R5$	No	No	No	No	No	No
	$R6$	No	Yes	Yes	Yes	No	No
	$R7$	No	No	No	No	No	No
Challenges	$C1$	Yes	No	No	Yes	Yes	No
	$C2$	Yes	Yes	No	No	Yes	No
	$C3$	Yes	No	No	Yes	Yes	Yes
	$C4$	No	No	Yes	Yes	No	No
	$C5$	No	No	No	No	No	No
	$C6$	Yes	No	No	Yes	Yes	Yes
	$C7$	No	No	No	No	No	No

TABLE III: Comparison of HP2P-SIP architectures

as Security Issues and Trustworthiness of Peers ($C4$) and bandwidth overload ($C6$) still remain major challenges.

The interest of this table is to highlight the major challenges of the moment and so allow researchers to easily find problems to attack in order to find solutions.

2) Comparison of HP2P-SIP Overlays multi-domains: We consider TABLE II. The same observations made in TABLE I for the P2P-SIP Client and fault tolerance are valid here. Indeed, these two requirements are still unresolved in the literature. Additionally, we find that in the existing litterature, the problem of clutter of super nodes remains still not solved. This is due to the fact that for all the architectures, all nodes of domains pass by super nodes to reach the main domain or the others.

Otherwise, we note that the fluidity of bandwidth ($C6$) depends on the number of DHT used. With several DHTs, the band is less crowded. This is due to the fact that with several DHTs, in the routing table of a node, are only the nodes belonging to its own domain and not the others. So the length of the routing table is greatly reduced.

As the first table, this one highlights the main challenges faced in hierarchical architectures. So it helps the researcher to be able to move on the problems yet to be solved in HP2P-SIP.

Since these two previous tables allow us to see the outstanding problems in hierarchical architectures, we give in the next section a way to implement a hierarchical system while minimizing the overload of the bandwidth (which is a major problem).

IV. Cost lookup-based analysis

In this section, we conduct a theoretical study in order to find a formula that can give exact number of messages generated in each of the three types of architecture described above. Contrary to [11] that compares flat systems and hierarchical systems and show that hierarchisation is benefical, in our case, we compare the hierarchical systems between them. As proved in the work of Bryan et *al.* in [25] and those of Zoels et *al.* in [7], the number of generated messages depends strongly on the location method used.

Indeed in [25], authors have given the number of messages generated (generically) by a node, according to the research method used. In [7], Zoels et *al.* treated the cost of research and maintenance only in a single domain architecture (one-layered overlay). For the lookup cost, they differentiate between lookup cost of super nodes (SN) and lookup cost of ordinary nodes (ON).

However, theses papers[7], [25] do not show what architecture consumes more the bandwidth than the other. Therefore, we conduct a theoretical study in order to find a formula that can give exact number of messages generated in each of the three types of architectures.

Contrary to [7], that use the number of exchanged messages by each node without explaining its real value, in our study, we formally determine this latter according to [25]. After this, we use the obtained expression in order to fine a general formula giving the total number of generated messages in each architecture. Finally, in each architecture, we formally determine the cost of each routing method, in terms of total number of exchanged messages. Note that in our case, we focus only on the number of generated messages by the lookup nodes. To do this, we consider Chord like underling DHT.

A. Hierarchical Single Domain Architecture (H-SDA)

Recall that this architecture is described in (figure *III-A1a* page 4). We will recall some formulas used in [7]. In these formulas, the terms which we are interested are $\mathcal{R}_{LKP}$ ones ($\mathcal{R}_{LKP,ON}$ and $\mathcal{R}_{LKP,SN}$). This later represent the number of sent/received messages by a node (Super Nodes (SN) or Ordinary Nodes (ON)) during the LookuP (LKP) process. Thus, based on [25], we calculate this terms. To do this, we use the same notation as used by Zoels et *al.* in [7] and adapting them according to our case. In the remainder of this section, we use the following notations.

- $\mathcal{N}_{SN}$: number of Super Nodes (SN).
- $\mathcal{N}_{ON}$: number of Ordinary Nodes (ON).
- $\mathcal{N}_{ON_i}$: number of ON attached to a SN i.

- $\mathcal{C}_{LKP}^{H-SDA}$: total cost for lookup traffic in H-SDA.
- $\mathcal{C}_{LKP,ON}^{H-SDA}$: lookup cost for all ordinary nodes in H-SDA.
- $\mathcal{C}_{LKP,SN}^{H-SDA}$: lookup cost for all super nodes in H-SDA.

Accordind to [7], the lookup cost in H-SDA is obtained as follows :

$$\mathcal{C}_{LKP}^{H-SDA} = \mathcal{C}_{LKP,ON}^{H-SDA} + \mathcal{C}_{LKP,SN}^{H-SDA} \quad (2)$$

In the following, we will determine respectively the ordinary node's and super node's lookup cost.

1) Lookup for ordinary nodes: To determine the ordinary node's lookup cost, we use the following notation.

- $\mathcal{R}_{LKP,ONj}$: number of sent/received messages by an ordinary node j.
- $\mathcal{C}_{ONj}^{s}$: cost generated by ordinary node j for sending message.
- $\mathcal{C}_{ONj}^{r}$: cost generated by ordinary node j for receiving message.

According to [7], the lookup cost for all ordinary node is given by the following equation :

$$\mathcal{C}_{LKP,ON}^{H-SDA} = \sum_{j=1}^{\mathcal{N}_{ON}} (\mathcal{C}_{ON_j}^{s} + \mathcal{C}_{ON_j}^{r}) \times \mathcal{R}_{LKP,ON_j} \quad (3)$$

2) Lookup for Super Nodes: To determine the ordinary node's lookup cost, we use the following notation.

- $\mathcal{R}_{LKP,SN_i}$: number of sent/received messages by super node i.
- $\mathcal{N}_{mess}^{H-SDA}$: number of sent/received messages by all SN in H-SDA.
- $\mathcal{C}_{SN_i}^{s}$: cost generated by super node i for sending message.
- $\mathcal{C}_{SN_i}^{r}$: cost generated by super node i for receiving message.

In [7], Zoels et *al.* show that the total number of generated message for each super node i (SN_i) and each ordinary node j (ON_j), that performs the lookups process in a chord ring is obtained as follows:

$$\mathcal{M}_{(SN_i)} = \mathcal{R}_{LKP,SN_i} \times \log_2(\mathcal{N}_{SN}) \quad (4)$$

$$\mathcal{M}_{(ON_j)} = \mathcal{R}_{LKP,ON_j} \times [\log_2(\mathcal{N}_{SN}) + 1] \quad (5)$$

Therefore, using equations 4 and 5, we calculate the total number of generated messages (noted $\mathcal{N}_{mess}^{H-SDA}$) by all nodes in H-SDA as follows:

$$\mathcal{N}_{mess}^{H-SDA} = \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{M}_{(SN_i)} + \sum_{i=1}^{\mathcal{N}_{SN}} \sum_{j=1}^{\mathcal{N}_{ON_i}} \mathcal{M}_{(ON_j)} \quad (6)$$

$$\mathcal{N}_{mess}^{H-SDA} = \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{LKP,SN_i} \times \log_2(\mathcal{N}_{SN}) + \sum_{i=1}^{\mathcal{N}_{SN}} \sum_{j=1}^{\mathcal{N}_{ON_i}} \mathcal{R}_{LKP,ON_j} \times [\log_2(\mathcal{N}_{SN}) + 1]$$

Finally, after reorganization, the total number of generated messages in H-SDA is obtained by the following formula:

$$\mathcal{N}_{mess}^{H-SDA} = \log_2(\mathcal{N}_{SN}) \times \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{LKP,SN_i} + [\log_2(\mathcal{N}_{SN}) + 1] \times \sum_{i=1}^{\mathcal{N}_{SN}} \sum_{j=1}^{\mathcal{N}_{ON_i}} \mathcal{R}_{LKP,ON_j} \quad (7)$$

As the number of exchanged messages (sent and received) depends on the lookup method used [25], after having determined a generic formula (Equation 7), we will focus on the cost of each lookup method. To dos this, we consider the following lookup methods:

- iterative or exhaustive-iterative
- full-recursive or source-routing-recursive (recursive)
- semi-recursive

According to [10], the exhaustive-iterative method consumes more bandwidth than the iterative method. Since in our case we are interested in minimizing the occupation of bandwidth, we will work with the iterative method. Similarly, the full-recursive methods and source-routing-recursive are the same. The only difference is that for full-recursive, the response is routed back to the originator recursively while for source routing-recursive the response is routed back to the originator along the reverse path of the routed messages [26]. According to the study in [10], source-routing-recursive method is less expensive than full-recursive method. Therefore in what follows, we will work with iterative methods, recursive and semi-recursive. In remain of this paper, we use the following notations:

- Ite=iterative
- Rec=recursive
- SRec=semi-recursive

a) Iterative lookup: In [25], Bryan et *al.* show the that, in the iterative lookup method, the number of generated messages by each node is $2 \times (n-1)$, where n is the number of nodes participating to the distribution. Thus, we apply this in different $\mathcal{R}_{LKP}$ and we obtain:

$$\mathcal{R}_{LKP,SN_i} = 2(\mathcal{N}_{SN} - 1) \quad (8)$$

$$\mathcal{R}_{LKP,ON_j} = 2(\mathcal{N}_{SN} - 1) + 2 \quad (9)$$

Note that:

$$\mathcal{R}_{LKP,ON_j} = 2\mathcal{N}_{SN} \quad (10)$$

Remark 4.1: For ordinary nodes, as the research is done by the super nodes, then we have $2(\mathcal{N}_{SN}-1)$ [7]. In addition, for an ordinary node, it is necessary to consider the message sends to its super node and the message it receives from it in case of response. Thus, we must add $+2$.

For all super nodes, the total number of generated messages during the lookup process is:

$$\sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{LKP,SN_i} = \sum_{i=1}^{\mathcal{N}_{SN}} 2(\mathcal{N}_{SN} - 1)$$

Thus, we obtain :

$$\sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{LKP,SN_i} = 2 \times \mathcal{N}_{SN} \times (\mathcal{N}_{SN} - 1) \quad (11)$$

For all ordinary nodes, the total number of generated messages during the lookup process is:

$$\sum_{j=1}^{\mathcal{N}_{ON_i}} \mathcal{R}_{LKP,ON_j} = \sum_{j=1}^{\mathcal{N}_{ON_i}} 2\mathcal{N}_{SN}$$

Thus, we obtain :

$$\sum_{j=1}^{\mathcal{N}_{ONi}} \mathcal{R}_{LKP,ON_j} = 2 \times \mathcal{N}_{ON_i} \times \mathcal{N}_{SN} \quad (12)$$

By injecting equations 11 and 12 in the equation 7, we deduce the following formula that give the total number of generated messages (noted $\mathcal{N}_{mess}^{H-SDA}(Ite)$).

$$\mathcal{N}_{mess}^{H-SDA}(Ite) = \log_2(\mathcal{N}_{SN}) \times 2 \times \mathcal{N}_{SN} \times (\mathcal{N}_{SN} - 1) + [\log_2(\mathcal{N}_{SN}) + 1] \times \sum_{i=1}^{\mathcal{N}_{SN}} 2 \times \mathcal{N}_{ON_i} \times \mathcal{N}_{SN}$$

Finally, after reorganization and simplification, in the iterative method, the total number of generated messages is obtained by the following formula :

$$\mathcal{N}_{mess}^{H-SDA}(Ite) = 2 \times \mathcal{N}_{SN} \times [(\mathcal{N}_{SN} - 1) \times \log_2(\mathcal{N}_{SN}) + [\log_2(\mathcal{N}_{SN}) + 1] \times \sum_{i=1}^{\mathcal{N}_{SN}} \times \mathcal{N}_{ON_i}] \quad (13)$$

b) Recursive lookup: In recursive lookup method, as the number of generated messages by a node is $2 \times (n-1)$ [25], therefore the total number of messages generated by all nodes in recursive lookup method (noted $\mathcal{N}_{mess}^{H-SDA}(Rec)$) is the same of iterative method. Thus

$$\mathcal{N}_{mess}^{H-SDA}(Ite) = \mathcal{N}_{mess}^{H-SDA}(Rec) = \mathcal{N}_{mess}^{H-SDA}(Ite/Rec) \quad (14)$$

c) Semi-Recursive lookup: Contrary to iterative or recursive method, in semi-recursive lookup method, the number of generated messages by a node is n, with n the number of nodes particiting to the distribution [25]. We apply this in different $\mathcal{R}_{LKP}$ and we obtain:

$$\mathcal{R}_{LKP,SN_i} = \mathcal{N}_{SN} \quad (15)$$

$$\mathcal{R}_{LKP,ON_j} = \mathcal{N}_{SN} + 2 \quad (16)$$

Then, we obtain :

$$\sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{LKP,SN_i} = \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{N}_{SN} = (\mathcal{N}_{SN})^2 \quad (17)$$

$$\sum_{j=1}^{\mathcal{N}_{ON_i}} \mathcal{R}_{LKP,ON_j} = \sum_{j=1}^{\mathcal{N}_{ON_i}} (\mathcal{N}_{SN} + 2) = \mathcal{N}_{ONi} \times (\mathcal{N}_{SN} + 2) \quad (18)$$

By injecting equations 17 and 18 in the equation 7, we deduce the following formula that give the total number of generated messages (noted $\mathcal{N}_{mess}^{H-SDA}(SRec)$) in semi-recursive method.

$$\mathcal{N}_{mess}^{H-SDA}(SRec) = (\mathcal{N}_{SN})^2 \times \log_2(\mathcal{N}_{SN}) + (\mathcal{N}_{SN} + 2) \times [\log_2(\mathcal{N}_{SN}) + 1] \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{N}_{ON_i} \quad (19)$$

Remark 4.2: In the particular case that we have the same number of ordinary nodes in each super node (*i.e.* $\mathcal{N}_{ON_i} = \mathcal{N}_{ON/SN}$), we have:

$$\sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{N}_{ON_i} = \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{N}_{ON/SN} = \mathcal{N}_{SN} \times \mathcal{N}_{ON/SN} \quad (20)$$

Therefore:

- In the iterative or recursive method, the number of exchanged messages is:

$$\mathcal{N}_{mess}^{H-SDA}(Ite/Rec) = 2 \times \mathcal{N}_{SN} \times [(\mathcal{N}_{SN} - 1) \times (\log_2(\mathcal{N}_{SN})) + \mathcal{N}_{SN} \times \mathcal{N}_{ON/SN} \times (\log_2(\mathcal{N}_{SN}) + 1)] \quad (21)$$

- In the semi-recursive method, the number of generated messages is:

$$\mathcal{N}_{mess}^{H-SDA}(SRec) = (\mathcal{N}_{SN})^2 \times \log_2(\mathcal{N}_{SN}) + \mathcal{N}_{SN} \times \mathcal{N}_{ON/SN} \times (\mathcal{N}_{SN} + 2) \times [\log_2(\mathcal{N}_{SN}) + 1] \quad (22)$$

B. Hierarchical Multi-Layered Architecture (H-MLA)

Recall that this architecture is described in (figure *III-A1b* page 4). In our case, sub-levels as top level use CHORD as DHT. In this the remain of this section, we use the following notations:

- K: number of sub levels.
- $\mathcal{N}_{P_i}$: number of nodes in sub-level i.
- $\mathcal{N}_{SN}$: number of super nodes in top level.
- $\mathcal{R}_{SN_i}$: number of sent/received messages by a super node i.
- $\mathcal{R}_{P_i}$: number of sent/received messages by a node i.

In H-MLA architecture, as all nodes participate in the distribution in their own level, thus if apply the formulas of Zoels et *al.* [7], we obtain:

- The number of messages (sent/received) generated (noted $\mathcal{M}_{(SN_i)}$) by a super node i that performs the lookup is :

$$\mathcal{M}_{(SN_i)} = \mathcal{R}_{SN_i} \times \log_2(\mathcal{N}_{SN}) \quad (23)$$

- The number of messages (sent/received) generated (noted $\mathcal{M}_{(P_j)}$) by a node j belonging to sub-level i that performs the lookup is :

$$\mathcal{M}_{(P_j)} = \mathcal{R}_{P_j} \times \log_2(\mathcal{N}_{P_i}) \quad (24)$$

Thus, the total number of messages sent and received in level i (noted $\mathcal{N}_{mess/i}^{H-MLA}$) and in Top Level ($TL$) (noted $\mathcal{N}_{mess/TL}^{H-MLA}$) are respectively given by the following equations:

- Total number of messages sent and received in Top level:

$$\mathcal{N}_{mess/TL}^{H-MLA} = \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{M}_{(SN_i)}$$

Therefore:

$$\mathcal{N}_{mess/TL}^{H-MLA} = \log_2(\mathcal{N}_{SN}) \times \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{SN_i} \quad (25)$$

- Total number of messages sent and received in level i:

$$\mathcal{N}_{mess/i}^{H-MLA} = \sum_{j=1}^{\mathcal{N}_{P_i}} \mathcal{M}_{(P_j)}$$

$$\mathcal{N}_{mess/i}^{H-MLA} = \log_2(\mathcal{N}_{P_i}) \times \sum_{j=1}^{\mathcal{N}_{P_i}} \mathcal{R}_{P_j} \quad (26)$$

According to equations 25 and 26, the total number of generated messages in H-MLA is:

$$\mathcal{N}_{mess}^{H-MLA} = \mathcal{N}_{mess/TL}^{H-MLA} + \sum_{i=1}^{k} \mathcal{N}_{mess/i}^{H-MLA} \quad (27)$$

Finally, after reorganization and simplification, we obtain the following equation:

$$\mathcal{N}_{mess}^{H-MLA} = \log_2(\mathcal{N}_{SN}) \times \sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{SN_i} + \sum_{i=1}^{k} (\log_2(\mathcal{N}_{P_i})) \times \sum_{j=1}^{\mathcal{N}_{P_i}} \mathcal{R}_{P_j} \quad (28)$$

In the following, we determine the total number of generated messages for each method (*iterative, recursive and semi-recursive)* in H-MLA.

1) For iterative and recursive methods: In iterative and recursive methods, the number of generated messages by a node is [25]:

- in each sub-level i, as all nodes ($\mathcal{N}_{P_i}$ in number) participate in the distribution, we have:

$$\mathcal{R}_{P_j} = 2 \times (\mathcal{N}_{P_i} - 1) \quad (29)$$

- in the top level, as all nodes ($\mathcal{N}_{SN}$ in number) participate in the distribution, we have:

$$\mathcal{R}_{SN_i} = 2 \times (\mathcal{N}_{SN} - 1) \quad (30)$$

Thus, the total number of generated messages by all nodes is:

- in each sub-level i :

$$\sum_{j=1}^{\mathcal{N}_{P_i}} \mathcal{R}_{P_j} = 2 \times \mathcal{N}_{P_i} \times (\mathcal{N}_{P_i} - 1) \quad (31)$$

- in the top level :

$$\sum_{i=1}^{\mathcal{N}_{SN}} \mathcal{R}_{SN_i} = 2 \times \mathcal{N}_{SN} \times (\mathcal{N}_{SN} - 1) \quad (32)$$

Therefore:

- the total number of generated messages in the top level is :

$$\mathcal{N}_{mess/TL}^{H-MLA}(Ite/Rec) = 2 \times \mathcal{N}_{SN} \times (\mathcal{N}_{SN} - 1) \times (\log_2(\mathcal{N}_{SN})) \quad (33)$$

- the number of messages generated by the set of sub-levels is:

$$\sum_{i=1}^{K} \mathcal{N}_{mess/i}^{H-MLA}(Ite/Rec) = 2 \times \sum_{i=1}^{K} \mathcal{N}_{P_i} \times (\mathcal{N}_{P_i} - 1) \times (\log_2(\mathcal{N}_{P_i})) \quad (34)$$

Using equations 33 and 34, we calculate the total number of generated messages in H-MLA by iterative or recursive method as follows:

$$\mathcal{N}_{mess}^{H-MLA}(Ite/Rec) = 2 \times \mathcal{N}_{SN} \times (\mathcal{N}_{SN} - 1) \times \log_2(\mathcal{N}_{SN}) + 2 \times \sum_{i=1}^{K} \mathcal{N}_{P_i} \times (\mathcal{N}_{P_i} - 1) \times \log_2(\mathcal{N}_{P_i}) \quad (35)$$

2) For semi-recursive method: According to [25] and by proceeding in the same way, we obtain:

- in each sub-level i :

$$\mathcal{R}_{P_j} = \mathcal{N}_{P_i} \quad (36)$$

- in the top level :

$$\mathcal{R}_{SN_i} = \mathcal{N}_{SN} \quad (37)$$

$$\mathcal{N}_{mess/TL}^{H-MLA}(SRec) = (\mathcal{N}_{SN})^2 \times (\log_2(\mathcal{N}_{SN})) \quad (38)$$

and

$$\sum_{i=1}^{K} \mathcal{N}_{mess/i}^{H-MLA}(SRec) = \sum_{i=1}^{K} (\mathcal{N}_{P_i})^2 \times (\ log_2(\mathcal{N}_{P_i})) \quad (39)$$

Therefore, we deduce the total number of generated messages in H-MLA by the semi-recursive method as follows:

$$\mathcal{N}_{mess}^{H-MLA}(SRec) = (\mathcal{N}_{SN})^2 \times (\log_2(\mathcal{N}_{SN})) + \sum_{i=1}^{K} (\mathcal{N}_{P_i})^2 \times (\log_2(\mathcal{N}_{P_i})) \quad (40)$$

Remark 4.3: In the particular case where we have, in all sub-levels, the same number of nodes (noted N_P), we'll have:

- In iterative and Recursive methods:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\sum_{i=1}^{K} \mathcal{N}_{mess/i}^{H-MLA}(Ite/Rec) = 2 \times K \times \mathcal{N}_P \times (\mathcal{N}_P - 1) \times (\log_2(\mathcal{N}_P)) \quad (41)$$

The total number of generated messages is :

$$\mathcal{N}_{mess}^{H-MLA}(Ite/Rec) = 2{\times}\mathcal{N}_{SN}{\times}(\mathcal{N}_{SN}{-}1){\times}(\log_2(\mathcal{N}_{SN})) + 2 \times K \times \mathcal{N}_P \times (\mathcal{N}_P - 1) \times (\log_2(\mathcal{N}_P)) \quad (42)$$

- In Semi-Recursive method:

$$\sum_{i=1}^{K} \mathcal{N}_{mess/i}^{H-MLA}(SRec) = K \times \mathcal{N}_P^2 \times (\log_2(\mathcal{N}_P)) \quad (43)$$

The total number of generated messages is :

$$\mathcal{N}_{mess}^{H-MLA}(SRec) = \mathcal{N}_{SN}^2 \times (\log_2(\mathcal{N}_{SN})) + K \times \mathcal{N}_P^2 \times (log_2(\mathcal{N}_P)) \quad (44)$$

Remark 4.4: As a hop in the overlay network corresponds to several hops in the physical network, we can minimize overloading the bandwidth of a physical network by acting on the structure of the overlay network. Indeed, if we limit the number of messages in the overlay network, it is the same in the physical network.

For example, assume that we have a physical network of N nodes on which we want to create an overlay network with 5 super nodes ($\mathcal{N}_{SN} = 5$). Moreover, assume that each of sub layers have 150 nodes each ($\mathcal{N}_p = 150$). We want that in the overlay network, the number of messages (for lookups) does not exceed 1000000.

Question : In how many levels does it organize the overlay network so that the maximum number of exchanged messages (sent/received) is equal to 1000000 ?

Answers:

- According the equation 42, if it is iterative or recursive method that is implemented, we will have $K = 4$ levels.
- According to the equation 44, if it is the semi-recursive method that is implemented, the number of levels is $K = 7$ levels.

whether

- $\mathcal{N}_{ij}$ the number of hops between physical nodes i and j of the physical network
- $\mathcal{N}_{maxHops}$ the maximum of $\mathcal{N}_{ij}$ whatever i and j belonging to the physical network, $\mathcal{N}_{maxHops} = Max(\mathcal{N}_{ij})$

So, in the physical network, the number of messages will be at most $1000000 \times \mathcal{N}_{maxHops}$

Thus, we see that by limiting the number of messages on the overlay network (eg by acting on the number of layers), we can limit the number of messages on the physical network.

In the table below (TABLE IV), we show the number of K levels required for an overlay network which the maximum number of messages is fixed (example with $\mathcal{N}_{SN} = 5$ and $\mathcal{N}_p = 150$).

Maximum number of messages	Ite/Rec	SRec
1000000	$k = 4$	$k = 7$
2000000	$k = 7$	$k = 13$
3000000	$k = 10$	$k = 19$
5000000	$k = 16$	$k = 31$

TABLE IV: The number of levels according to the number of messages

C. Hierarchical Multi-Domain Architecture (H-MDA)

Recall that this architecture is described in (figure *III-A2a* page 5)

In our case, we use Chord as DHT in both the main domain and in subdomains.

- K: number of sub domains.
- $\mathcal{N}_{P_i}$: number of nodes in sub-domain i.
- $\mathcal{N}_{SN}$: number of super nodes in main domain.

As we proceeded with previous architectures, based on the same assumptions, we determine the number of exchanged messages as follows.

$$\mathcal{N}_{mess}^{H-MDA} = \mathcal{N}_{mess/TL}^{H-MDA} + \sum_{i=1}^{k} \mathcal{N}_{mess/i}^{H-MDA} \quad (45)$$

After applying to different methods, we obtain:

- in iterative or recursive method:

$$\mathcal{N}_{mess}^{H-MDA}(Ite/Rec) = 2{\times}\mathcal{N}_{SN}{\times}(\mathcal{N}_{SN}{-}1){\times}(\log_2(\mathcal{N}_{SN})) + 2 \times K \times \mathcal{N}_P \times (\mathcal{N}_P - 1) \times (\log_2(\mathcal{N}_P)) \quad (46)$$

- in semi-recursive method:

$$\mathcal{N}_{mess}^{H-MDA}(SRec) = (\mathcal{N}_{SN})^2 \times (\log_2(\mathcal{N}_{SN})) + K \times (\mathcal{N}_P)^2 \times (\log_2(\mathcal{N}_P)) \quad (47)$$

Similarly to H-MLA, this allows us, when we want to set up an overlay network of the type H-MDA, to know how many sub-domains must be composed the network in order to not clutter the bandwidth.

D. Comparative study of the number of messages generated depending on the architecture and the method of localization

In this section, we give a summary table of the number of messages generated by the different architectures based on the research methods used. Our goal is to observe the evolution of the number of messages generated as and as we vary the number of super nodes in the overlay network. The number of super nodes is arbitrarily selected in accordance with the same pitch.

Assumptions:

We start from a total number of nodes equal in different architectures. We assume to have a network of 1000 nodes(i.e $\mathcal{N} = 1000$).

We choose arbitrarily the number of super nodes (maintaining the same pitch) and observe the evolution of the number of generated messages.

$\mathcal{N}_{SN}$=number of super nodes

$\mathcal{N}_{ON/SN}$=number of ordinary nodes per super node

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE V: Comparison between hierarchical P2P-SIP architectures according to routing method used

		Number of generated messages			
		H-SDA		H-MLA / H-MDA	
		Ite/Rec	SRec	Ite/Rec	SRec
Number of super nodes	40	502128, 5	263415, 2	219076, 4	114152, 6
	80	1204557, 3	616842, 5	152784, 3	80210, 4
	120	2019008, 4	1025514, 2	234220, 8	120579, 2
	160	2929036, 0	1481667, 6	394253, 4	200469, 6
	200	3374484, 9	1702601, 4	618050, 9	312154, 3
	240	5011373, 7	2524685, 7	916198, 5	461516, 9
	280	5564533, 9	2799880, 4	1274683, 9	640759, 3
	320	7426397, 5	3733759, 9	1703322, 3	855403, 5
	360	7115550, 8	3574500, 7	2196254, 2	1101824, 2
	400	8931186, 9	4484481, 1	2760318, 9	1384216, 9

K = number of levels other than the top
$\mathcal{N}_P$= number of nodes in each level K

- In H-SDA:
  $\mathcal{N} = \mathcal{N}_{SN} + \mathcal{N}_{SN} \times \mathcal{N}_{ON/SN}$
  By choosing arbitrarily the values of $\mathcal{N}_{SN}$, we calculate $\mathcal{N}_{ON/SN}$ and replace in the equations of $\mathcal{N}_{mess}^{H-SDA}(Ite/Rec)$ and $\mathcal{N}_{mess}^{H-SDA}(SRec)$ (equation 21 and equation 22)

- In H-MLA or H-MDA:
  $\mathcal{N} = \mathcal{N}_{SN} + K \times \mathcal{N}_P$.
  K and $\mathcal{N}_P$ are inversely proportional, since we start with a fixed number of nodes (equal to 1000).
  By choosing arbitrarily the values of $\mathcal{N}_{SN}$, we calculate $\mathcal{N}_P$ in the same way we calculate $\mathcal{N}_{ON/SN}$. Then we calculate K and replace in the equations of $\mathcal{N}_{mess}^{H-MLA}(Ite/Rec)$ and $\mathcal{N}_{mess}^{H-MLA}(SRec)$ (equation 42 and equation 44). See that $\mathcal{N}_{mess}^{H-MLA}$ and $\mathcal{N}_{mess}^{H-MDA}$ have the same formula.

- The table (TABLE V) shows that whatever the number of super nodes and whatever the type of architecture, the iterative or recursive methods are more intensive in number of messages that the semi-recursive method. As the number of super nodes increases, the number of generated messages increases. This is due to the fact that the greater the number of super nodes increases, the architecture tends to a flat overlay network because we have a total number of fixed nodes. In other words, the number of ordinary nodes is reduced because some of them are selected to become supernodes. Therefore the number of nodes participating in the distribution (that is to say, the super nodes) becomes larger.
- Similarly, if we consider the recursive or iterative methods (columns 3 and 5 of TABLE V), the table shows that the H-SDA architecture generates more messages than architectures H-MLA or H-MDA. It is the same when we consider the semi-recursive method (columns 4 and 6 of TABLE V). Indeed, in H-SDA, only the super nodes participate in the distribution. Now, in addition to their own messages, super nodes must support all messages of ordinary nodes that are attached to them. That's why when they become more numerous, the number of messages is growing faster in H-SDA than in H-MLA or H-MDA. Because, in H-MLA or H-MDA, for each node is taken into account only its own messages.
- These two points on the table show that by using the semi-recursive method with an architecture H-MLA or H-MDA (ie multilevel) research will be more effective.

V. Conclusion and Open Issues

In this paper, we have shown the behavior of the various hierarchical P2P-SIP architectures according to the requirements of P2P-SIP networks. This allowed us to highlight the many existing challenges in hierarchical architectures. We have also determined the number of messages that can be generated in a HP2P-SIP overlay network, and this, whatever the location method used. Through this last part, we have set up a way to design a hierarchical overlay system by limiting bandwidth overload in the context of lookup nodes. We also compared in terms of generated messages between different hierarchical architectures but also between routing methods.
In the case of our outlook, we will conduct an experimental study by simulations with OverSim simulator for evaluating the average performance of the different methods in terms of communication costs, occupancy of bandwidth, delays, and so on.

References

[1] S. Yahiaoui, Y. Belhoul, N. Nouali-Taboudjemat, and H. Kheddouci, "Adsip: Decentralized sip for mobile ad hoc networks," in *Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on*, March 2012, pp. 490–495.
[2] X. Zheng and V. Oleshchuk, "A Survey on Peer-to-Peer SIP Based Communication Systems," *Peer-to-Peer Networking and Applications*, vol. 3, no. 4, pp. 257–264, 2010.
[3] I. Martinez-Yelmo, C. Guerrero, R. Cuevas, and A. Mauthe, "A hierarchical p2psip architecture to support skype-like services," in *Proceedings of the 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing*, 2009, pp. 316–322.
[4] H. Ma, B. Xu, H. Wan, and C. Li, "A Hierarchical P2P Architecture for SIP Communication," in *Proceedings of the International Conference on Next Generation Mobile Applications, Services and Technologies*, ser. NGMAST '07, 2007, pp. 130–135.
[5] J. Shi, Y. Wang, L. Gu, L. Li, W. Lin, Y. Li, Y. Ji, and P. Zhang, "A hierarchical peer-to-peer sip system for heterogeneous overlays interworking," in *Proceedings of the IEEE Global Telecommunications Conference*, ser. GLOBECOM '07, 2007, pp. 93–97.
[6] M. S. Islam, S. A. Rahman, R. Ahmen, and M. H. Raju, "A hierarchical overlay design for peer to peer and sip integration," *International Journal of Computer Science and Information Security*, vol. 9, no. 6, pp. 94–99, 2011.

[7] S. Zoels, Z. Despotovic, and W. Kellerer, "Cost-based analysis of hierarchical dht design," in *Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing*, ser. P2P '06, 2006, pp. 233–239.
[8] O. Bravo, A. Costa, and M. Nicolau, "Design and implementation of a hierarchical sip-based peer-to-peer network," in *Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on*, Sept 2012, pp. 1–9.
[9] L. Garcece-Erice, E. Biersack, P. Felber, K. Ross, and G. Urvoy-Keller, "Hierarchical peer-to-peer systems," ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2790, pp. 1230–1239.
[10] I. Baumgart and B. Heep, "Fast but economical: A simulative comparison of structured peer-to-peer systems," in *Next Generation Internet (NGI), 2012 8th EURO-NGI Conference on*. IEEE, 2012, pp. 87–94.
[11] M. Artigas, P. Lopez, and A. Skarmeta, "A comparative study of hierarchical dht systems," in *Local Computer Networks, 2007. LCN 2007. 32nd IEEE Conference on*, Oct 2007, pp. 325–333.
[12] W. Liu, J. Song, and J. Yu, "An overlapping structured {P2P} for {REIK} overlay network," *Physics Procedia*, vol. 33, no. 0, pp. 1022 – 1028, 2012, 2012 International Conference on Medical Physics and Biomedical Engineering (ICMPBE2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1875389212014812
[13] R. Khan and R. Hasan, "Secp2psip: A distributed overlay architecture for secure p2psip," 2014.
[14] A. Anitha, J. JayaKumari, and G. Mini, "A survey of p2p overlays in various networks," in *Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on*, July 2011, pp. 277–281.
[15] H. Yongfeng, S. Tang, and Y. Yip, "A new security architecture for sip based p2p computer networks," *Journal of Computer Science, Informatics and Electrical Engineering*, vol. 2, no. 1, 2008.
[16] T. Koskela, O. Kassinen, J. Korhonen, Z. Ou, and M. Ylianttila, "Peer-to-peer Community Management Using Structured Overlay Networks," in *Proceedings of the International Conference on Mobile Technology, Applications, and Systems*, ser. Mobility '08, 2008, pp. 10:1–10:6.
[17] J. Furness, M. Kolberg, and M. Fayed, "An evaluation of chord and pastry models in oversim," in *Modelling Symposium (EMS), 2013 European*, Nov 2013, pp. 509–513.
[18] K. Singh and H. Schulzrinne, "Peer-to-peer Internet Telephony Using SIP," in *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV '05, 2005, pp. 63–68.
[19] H.-W. Ferng and I. Christanto, "A globally overlaid hierarchical p2p-sip architecture with route optimization," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 11, pp. 1826–1833, Nov 2011.
[20] D. A. Bryan and B. B. Lowekamp, "Sosimple: A serverless, standards-based," in *P2P SIP Communication System, Proc. of the International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications*. IEEE Press, 2005, pp. 42–49.
[21] K. Singh and H. Schulzrinne, "Peer-to-peer internet telephony using sip," in *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV '05. New York, NY, USA: ACM, 2005, pp. 63–68.
[22] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications," *ACM SIGCOMM Computer Communication Review,*, vol. 31, no. 4, pp. 149–160, 2001.
[23] I. Diane, I. Niang, and B. Gueye, "A hierarchical dht for fault tolerant management in p2p-sip networks," in *Proceedings of the 2010 IEEE 16th International Conference on Parallel and Distributed Systems*, ser. ICPADS '10, 2010, pp. 788–793.
[24] L. Le and G.-S. Kuo, "Hierarchical and breathing peer-to-peer sip system," in *Proceedings of the IEEE International Conference on Communications,*, ser. ICC '07, 2007, pp. 1887–1892.
[25] D. Bryan, B. Lowekamp, and M. Zangrilli, "The design of a versatile, secure p2psip communications architecture for the public internet," in *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, April 2008, pp. 1–8.
[26] "OverSim The Overlay Simulation Framework." http://www.oversim.org/wiki/OverSimKbrRouting.

Abel DIATTA is currently a PhD student in Computer Science at the University Cheikh Anta Diop in Dakar (Senegal). His research concern: (i) the P2P-SIP networks, fault tolerance, security in P2P-SIP networks; (ii) improving the quality of the bandwidth; (iii) the hierarchical DHT formatting and routing protocols; (iv) the mobility, the Internet of Things in P2P networks, etc.

Ibrahima NIANG is an Associate Professor in Computer Science at the Faculty of Sciences and Technology of the Universit Cheikh Anta Diop de Dakar (UCAD), Senegal. He received the Ph.D. degree in computer science in 2002 at Paris V University. His research activities concern QoS, mobility and security in wireless networks and distributed systems (P2PSIP). In recent years, he worked on research and development related to water management using biosensor networks.

Mandicou BA is currently a postdoctoral researcher at CReSTIC laboratory of the University of Reims Champagne-Ardenne, France. He received his Ph.D in computer Science on 2014 from the University of Reims Champagne-Ardenne. He received his Master degree on Networking and Telecommunication from the University Cheikh Anta Diop (UCAD), Dakar, Senegal, in October 2010. His research interests include P2P networks, ad hoc and wireless sensor networks, csloud computing, Internet of Things, self-stabilization, clustering, energy optimization, routing protocols, aggregation, security, performance evaluation and simulation.

Unweighted Class Specific Soft Voting based ensemble of Extreme Learning Machine and its variant.

Sanyam Shukla
CSE department, M.A.N.I.T,
Bhopal, India

R. N. Yadav
ECE department, M.A.N.I.T,
Bhopal, India

***Abstract*— Extreme Learning Machine is a fast real valued single layer feed forward neural network. Its performance fluctuates due to random initialization of weights between input and hidden layer. Voting based Extreme Learning Machine, VELM is a simple majority voting based ensemble of Extreme learning machine which was recently proposed to reduce this performance variation in Extreme Learning Machine. A recently proposed class specific soft voting based Extreme Learning Machine. CSSV-ELM further refines the performance of VELM using class specific soft voting. CSSV-ELM computes the weights assigned to each class of component classifiers using convex optimization technique. It assigns different weights assuming different classifiers perform differently for different classes. This work proposes Un-weighted Class Specific Soft Voting based ensemble, UCSSV-ELM a variants of CSSV-ELM. The proposed variant uses class level soft voting with equal weights assigned to each class of component classifiers. Here all the classifiers are assumed to be equally important. Soft voting is used with the classifiers that have probabilistic outputs. This work evaluates the performance of proposed ensemble using both ELM and a variant of ELM as base classifier. This variant of ELM differs from ELM as it uses sigmoid activation function at output layer to get probabilistic outcome for each class. The result shows that the Un-weighted class specific soft voting based ensemble performs better than majority voting based ensemble.**

Keywords—Ensemble Pruning; Extreme learning Machine; soft voting, probabilistic output.

I. INTRODUCTION

Most of the problems like intrusion detection, spam filtering, biometric recognition etc. are real valued classification problems. So many classifiers like SVM, C4.5, Naive Bayes etc. are available for real valued classification. Extreme learning machine, ELM [1] is a state of art classifier for real valued classification problems. ELM, is a feed forward neural network in which the weights between input and hidden layer are assigned randomly whereas, the weights between hidden and output layer are computed analytically. This makes extreme learning machine fast compared to other gradient based classifiers. But the random initialization of input layer weights leads to fluctuation in performance of extreme learning machine. Any change in training dataset or change in the parameters of the classification algorithm leads to performance fluctuation. This fluctuation in performance is known as error due to variance. Various Ensembling approaches like bagging[2], adboost.M1[3], adaboost.M2[3] have been designed to reduce this error due to variance and thereby increasing the performance of the base classifier. Various Variants of ELM [4]–[10] based on Ensembling techniques have been proposed to enhance the performance of ELM. This work also proposes a new Unweighted (equally weighted) Class Specific Soft Voting based classifier ensemble using ELM as base classifier. It also evaluates the proposed classifier using ELM variant as base classifier, which uses sigmoid activation function to get probabilistic outcome. This ELM variant was used in [7] as a base classifier with adaboost ensembling method to enhance the performance of ELM.

In the next section this paper discusses related work i.e. ELM, VELM and other various ELM based ensembles.. After this section, this paper describes the proposed work. After that, this paper describes the experimental setup and results obtained. The last section consists of conclusion and future work.

II. RELATED WORK

This section contains the brief review of the fundamental topics which were proposed earlier and are important from the perspective of the proposed work.

A. Extreme Learning Machine

ELM [1] is a fast learning Single Layer Feed Forward Neural Network. Let the input to ELM be N training samples with their targets $[(x_1, t_1), (x_2, t_2), \ldots, (x_j, t_j), \ldots, (x_N, t_N)]$ Here $j=1, 2, \ldots, N$, $x_j = [x_{j1}, x_{j2}, \ldots, x_{jF}]^T \in R^m$ and $t_i \in 1, 2, ..., C$. Here, F and C are the number of features and classes respectively. Fig. 1 shows the architecture of ELM.

In ELM, the number of input neurons is equal to number of input features. The number of hidden neurons, NHN is chosen as per complexity of the problem. The number of output neurons is equal to C. In ELM, the weights between input and hidden neurons are assigned randomly and weights between output and hidden neurons are computed analytically. This reduces the overhead of tuning the learning parameters, which makes it fast and more accurate compared to other gradient based techniques. In ELM, the neurons in the hidden layer use non-linear activation function while neurons in the output layer use linear activation function. The activation function of hidden layer neurons is any infinitely differentiable like Sigmoid function, Radial Basis function etc. Vector, $w_i = [w_{1i}, w_{2i}, \ldots, w_{Fi}]^T$ represents the weight vector connecting F^{th} input neurons to the i^{th} hidden neurons , where $i=1, 2, \ldots, NHN$, b_i

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the bias of i^{th} hidden neuron. The output of the i^{th} hidden neuron is represented by $g(w_i. x + b_i)$. Here, g is the activation function.

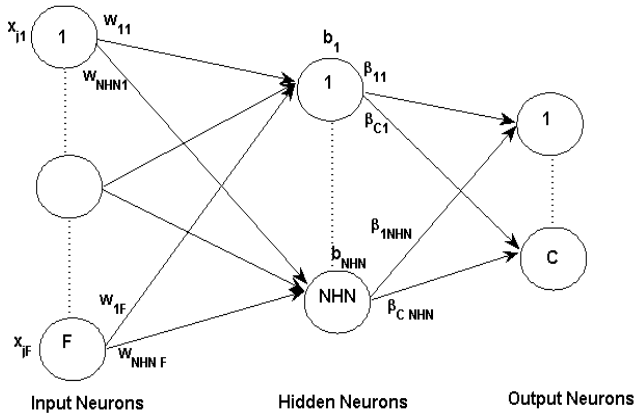


Fig 1. Architecture of ELM

The hidden layer output of all training samples can be represented as follows.

$$H_{train}(N\,X\,NHN)=\begin{bmatrix} g(w_1.x_1+b_1) & \cdots & g(w_{NHN}.x_1+b_{NHN}) \\ \vdots & \ldots & \vdots \\ g(w_1.x_N+b_1) & \cdots & g(w_{NHN}.x_N+b_{NHN}) \end{bmatrix}$$

Vector, $\beta_k = [\beta_{1k}, \beta_{2k}, ..., \beta_{NHN\ k}]^T$ represents the output weight vector which connects the k^{th} output neurons to the hidden neurons, here k = 1, 2, ..., C. The target of x_j i.e. t_j is represented as target vector $T_j = [T_{j1}, T_{j2}, \ldots, T_{jC}]^{T,}$ where t_{jk} = +1 if t_j = k else t_{jk} = -1. Vector, β_k is determined analytically using the equation given below.

$$\beta = (H_{train})^{+} T_{train}$$

$$Here, \quad \beta(NHN\,X\,C)=\left[\beta_1^{\,T}, \beta_2^{\,T}, \ldots\ldots\ldots, \beta_C^{\,T}\right] \text{ and}$$

$$T_{train}(NXC)=\begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}$$

H_{test} (n x NHN), is the hidden layer output matrix for the n testing instances. The predicted output for testing instances i.e. Y_{test} (n x C) is determined using the following equation

$$Y_{test} = H_{test}\,\beta$$

L_{test}(n x 1), output label of n testing instances is determined using this equation

$$L_{test} = \arg^{row}_{\max}(Y_{test})$$

The arg function returns the index of the maximum value for each row of Y_{test}.

B. Voting Based Extreme Learning Machine

As the parameters of ELM between the input and hidden layer are assigned random weights, the performance of ELM fluctuates. Due to this some instances which are close to bayes boundary are misclassified. V-ELM [4] solves this problem, by generating a number of classifiers, succeeded by voting for finding the prediction of the ensemble. Component classifiers of the ensemble are generated by randomly assigned learning parameters between input and hidden layer. Thus, each classifier is independent of the other. For n testing instances, the output label (L_{test}) for k classifiers is obtained where k = 1, 2…NCE. The final predicted output (FP_x) of x^{th} instance is obtained by the majority voting of component classifiers.

$$FP_x = \text{mode}(L_{test})$$

Here, $L_{test} = [L^1_{test}, L^2_{test}, \ldots, L^{NCE}_{test}]$

The mode operation calculates the class to which the maximum numbers of classifiers are voting. Taking an example of binary label where, the output is either positive or negative, NCE =30. Let for any test instance 18 classifiers give positive output whereas, 12 give negative output then the final output of V-ELM is positive.

C. Class Specific Soft Voting Based Extreme Learning Machine

Majority weighting can be categorised as simple voting, weighted voting and soft voting. In simple voting all the classifiers have equal weight. In weighted voting the classifier which performs better has more weight compared to the inferior classifiers. For both simple and weighted voting, the outcome of a classifier for any instance for a particular class is either 1 or 0 depending on whether it belongs to or does not belongs to that class repectively.

Class Specific Soft Voting Based Extreme Learning Machine, CSSV-ELM [5] improves the performance of VELM by employing soft voting along with ensemble pruning. The soft voting corresponds to the classifiers that have probabilistic outcome. Weighted soft computing method can be defined at three levels: classifier level, class level or instance level. CSSV-ELM is a classifier ensemble using soft voting at class level with base classifier as ELM. A classifier ensemble having T classifiers, employed to solve an m class classification problem, then soft voting based classifier ensemble will have

1) T different weights for classifier level weighted voting. Here each class of a particular classifier has equal weight.

2) T*m different weights one for each class of a particular classifier. Weighted voting is at the classifier level. Soft voting assigns weights to the classifier at class level. It can be used with classifiers with probabilistic outcomes. CSSV-ELM is improved VELM algorithm with soft voting. This work uses convex optimization for assigning different weights corresponding to the each class of the classifier.

D. Dynamic ensemble ELM based on sample entropy.

[7] enhances the performance of ELM by using adaboost ensembling method. To get the probabilistic output for each

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

class it uses a variant of ELM with sigmoidal activation function at the output layer. It generates the component classifiers of the ensemble using adaboost algorithm. The theme of this approach is to incorporate the confidence of prediction in final voting. Lower values of entropy indicate higher confidence of prediction. The component classifiers having normalized entropy lower than threshold participates in finding final prediction for a given test instance.

III. Proposed Work

The proposed classifier ensemble is similar to simple voting as it assigns all the classifier equal weights. It is different from simple voting as outcome of the classifier chosen as base classifier is treated as probabilistic output. This work is different from [5] as CSSV-ELM assigns class specific weight using convex optimization whereas the proposed work assigns equal weights to all the classes of any classifier. The combined outcome of the proposed classifier can be computed as

$$f_{com}(x_n) = \sum_{i=1}^{T} f_i(x_n)$$

$f_i(X_n)$ is a m dimensional vector representing the probabilistic output of i[th] classifier for all the m classes. The final outcome:

$$L_{x_n} = \arg\max_{class} \left(f_{com}(x_n)\right)$$

CSSV-ELM assigns different weights to all the T*m different classes. The hypothesis behind this is classifiers with better training accuracy will perform better for test data. So the classifiers with better training accuracy should be assigned more weights. This thing is not always true. The classifier with higher training accuracy might be having over fitting problem and assigning higher weight to such classifiers may degrade the performance. This fact is illustrated by the results shown in Table 1 and Fig. 1 for bupa dataset

Table I: Gmean of ELM for bupa dataset for NHN =100

Trial	Train Gmean	Test Gmean	Train Gmean - 0.620552	Test Gmean - 0.847221
1	0.622308	0.835457	0.001756	-0.01176436
2	0.615891	0.851214	-0.00466	0.003993451
3	0.617936	0.852095	-0.00262	0.004874101
4	0.638823	0.858269	0.018271	0.011047562
5	0.598993	0.839213	-0.02156	-0.00800811
6	0.61024	0.841011	-0.01031	-0.00620986
7	0.621662	0.855387	0.00111	0.008166466
8	0.630773	0.84175	0.010221	-0.00547127
9	0.641566	0.843424	0.021014	-0.00379737
10	0.607326	0.854389	-0.01323	0.007167761
Mean	0.620552	0.847221	-2.7E-07	0.226668837

The objective of this work is to compare Unweighted Class specific soft voting and Simple majority voting. This work also studies whether to use or not to use sigmoid activation for normalizing ELM output so as to get probabilistic output. Majority voting is one of the most common methods for making ensemble. This work compares simple majority voting and soft voting assuming all the classifiers are equally important. This work considers two different classifiers ELM [1] and a variant of ELM used in [7] as base classifier for constructing ensemble. This work constructs 4 different ensembles using the 2 above classifiers and the 2 voting schemes. Applying soft voting on ELM is different from [] as all the classifiers are assumed to be equally important.

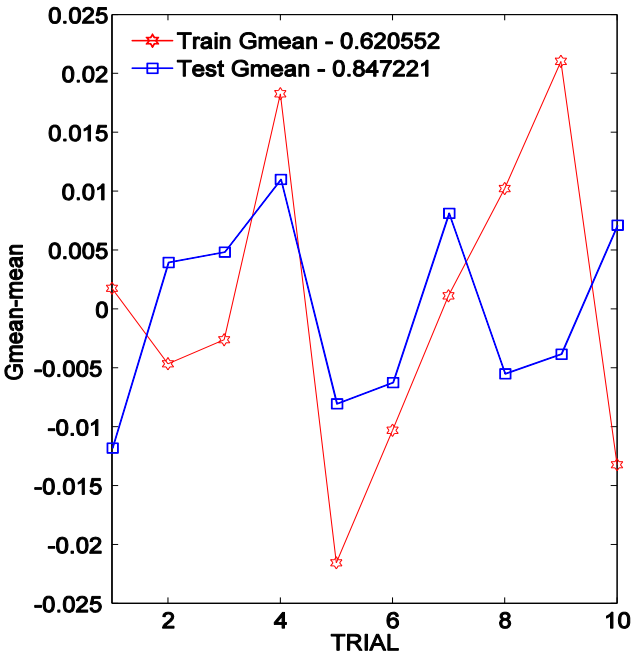


Fig. 2. Display of deviation from mean performance of ELM for bupa dataset for NHN=100

The algorithm of proposed classifier is as follows:

Algorithm for UCSSV-ELM

Input: N = Number of training instances, n = Number of testing instances, F = number of features, C = Number of class labels, Training Data, P_{train} (N x F), Training Target, t_{train}(N x 1), Processed Training Target, T_{train}(N x C), Testing Data, P_{test} (n x F) , Testing Target, t_{test} (n x 1)

Training Phase

I. Create k classifiers of the ensemble using following steps

for k = 1 : NCE

1. Take suitable number of NHN as per complexity of the problem.
2. Randomly generate the input weight of k[th] classifier, V^k (F X NHN) connecting the input and hidden neurons, and the bias, b^k (1 X NHN) of k[th]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

classifier for each hidden neuron.

3. Now calculate the bias matrix of k^{th} classifier for training data, B^k_{train} (N X NHN) by replicating b^k , N times.
4. Compute hidden neuron output matrix of k^{th} classifier, H^k_{train} (N X NHN) using the following equation.

$$H^k_{train} = g\left(\left(P_{train} * V^k\right) + B^k_{train}\right)$$

Here, g is sigmoid activation function.

5. Calculate the output weight of k^{th} classifier, β^k (NHN X C) by using the following equation.

$$\beta^k = (H^k_{train})^+ T_{train}$$

6. Compute the Predicted Output for training dataset, Y^k_{train}(N X C) using the learning parameters, β^k and Bias Matrix, H^k_{train} as follows.

For ELM

$$Y^k_{train} = (H^k_{train} * \beta^k)$$

For ELM variant with using sigmoid function to get probabilistic outcome

$$Y^k_{train} = Sigmoid(H^k_{train} * \beta^k)$$

7. Store the learning parameters NHN, V^k, b^k, β^k, Y^k_{train} .

End

II. Compute predicted label, L_{train} (N X 1) as follows:

$$SY_{train} = \sum_{k=1}^{NCE} Y^k_{train}$$

$$L_{train} = \arg^{row}_{\max}\left(SY_{train}\right)$$

Testing Phase

I. ***for*** k = 1 : NCE

1. Using the k^{th} bias (b^k), calculate the Bias Matrix of testing data, B^k_{test} (n X NHN) by replicating the bias, b^k n times.
2. Using the learning parameters of k^{th} classifier, V^k (F X NHN) and B^k_{test} (n X NHN). Compute hidden neuron output Matrix, H^k_{test} (n X NHN) using the following equation.

$$H^k_{test} = g\left(\left(P_{test} * V^k\right) + B^k_{test}\right)$$

3. Using the learning parameter of k^{th} classifier, β^k calculate the predicted output for testing dataset , Y^k_{test}(n x C) using the following equation

For ELM

$$Y^k_{test} = H^k_{test} * \beta^k$$

For ELM variant with using sigmoid function to get probabilistic outcome

$$Y^k_{test} = Sigmoid(H^k_{test} * \beta^k)$$

4. Store the computed output for test data (Y^k_{test}).

end

II. Compute predicted label, L_{test} (n X l) as follows:

$$SY_{test} = \sum_{k=1}^{NCE} Y^k_{test}$$

$$L_{test} = \arg^{row}_{\max}\left(SY_{test}\right)$$

III. Compute the testing overall accuracy and Gmean using L_{test} and actual target.

IV. EXPERIMENTAL SETUP

A. *Data Specification*

The proposed work is evaluated using 17 datasets, downloaded from the Keel-data set Repository [11]. The data sets in Keel Repository are available in 5 fold cross validation format i.e. for each dataset we have 5 training and testing sets. The specifications of testing and training datasets used for evaluation are shown in the Table II.

Table II. Specifications of datasets used in experimentation

DATA SET	Number of classes	Number of Attributes	Number of Training instances	Number of Testing instances
APPENDICITIS	2	7	84	22
BANANA	2	2	4240	1060
BUPA	2	6	276	69
CHESS	2	36	2556	640
GLASS0	2	9	171	43
HABERMAN	2	3	244	62
HAYES-ROTH	3	4	128	32
IONOSPHERE	2	33	280	71
IRISO	2	4	120	30
MONK-2	2	6	345	87
PHONEME	2	5	4323	1081
PIMA	2	8	614	154
RING	2	20	5920	1480
SA_HEART	2	9	369	93
SONAR	2	60	166	42
SPECTFHEART	2	44	213	54
TITANIC	2	3	1760	441

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table III: Average Overall Accuracy of Test Dataset for ELM and its variant

Base Classifier->	ELM				ELM variant using Sigmoid			
	UCSSV-ELM		VELM		UCSSV-ELM_S		VELM_S	
Dataset	NHN	AOA	NHN	AOA	NHN	AOA	NHN	AOA
appendicitis	10	87.71	10	87.99	10	**88.66**	10	88.18
banana	80	90.3	90	**90.31**	90	90.28	90	90.29
bupa	20	**73.04**	20	72.67	20	**73.04**	20	73.19
chess	90	95.37	100	95.39	100	**95.49**	100	95.4
glass0	60	**81.32**	60	79.58	40	79.45	60	79.86
haberman	10	73.53	10	**73.89**	20	73.86	10	73.72
hayes-roth	30	**75.63**	30	74.56	80	**75.63**	30	74.38
ionosphere	70	**92.89**	100	92.43	100	92.6	90	92.46
iris0	10	**100**	10	**100**	10	**100**	10	**100**
monk-2	70	**97.22**	70	**97.22**	30	95.35	30	94.65
phoneme	100	84.22	100	**84.33**	100	84.31	100	84.32
pima	20	77.6	20	77.54	20	**77.86**	20	77.6
ring	100	**94.64**	100	94.57	90	94.57	100	94.59
saheart	30	**73.8**	20	73.34	20	73.36	20	73.26
sonar	80	**86.49**	80	85.26	70	86.05	90	84.78
spectfheart	20	**79.4**	20	**79.4**	20	**79.4**	20	79.4
titanic	20	**78.87**	20	**78.87**	20	**78.87**	20	78.87

.

Table IV: Gmean of Test datasets for ELM and its varian

Base Classifier->	ELM				ELM variant using Sigmoid			
	UCSSV-ELM		VELM		UCSSV-ELM_S		VELM_S	
Dataset	**NHN**	**GMEAN**	**NHN**	**GMEAN**	**NHN**	**GMEAN**	**NHN**	**GMEAN**
appendicitis	10	68	10	69.63	10	**72.14**	10	70.07
banana	80	89.92	90	**89.94**	90	89.92	90	89.92
bupa	20	69.73	20	69.46	20	**69.96**	20	70.02
chess	90	95.31	100	95.34	100	**95.45**	100	95.35
glass0	60	**78.8**	100	77.98	90	78.69	60	77.71
haberman	20	48.88	20	48.87	20	**48.91**	20	49.29
hayes-roth	30	73.21	30	71.2	80	**76.48**	40	71.67
ionosphere	70	**90.26**	100	89.57	100	89.82	90	89.63
iris0	10	**100**	10	**100**	10	**100**	10	**100**
monk-2	70	**97.33**	70	**97.33**	30	88.71	20	87.31
phoneme	100	80.26	100	**80.5**	100	80.48	100	80.45
pima	100	70.77	90	70.18	40	**70.8**	20	70.51
ring	100	**94.44**	100	94.37	90	94.37	100	94.38
saheart	30	**64.99**	30	64.58	20	64.44	30	64.41
sonar	80	**86.16**	80	84.69	70	85.42	90	84.12
spectfheart	100	**41.44**	100	40.57	70	38.59	100	38.84
titanic	10	**67.08**	10	**67.08**	10	**67.08**	10	**67.08**

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table V. Result of wilcoxon signed rank test

	UCSSV-ELM	VELM	UCSSV-ELM_S
UCSSV-ELM			
VELM	**0.0676**		
UCSSV-ELM_S	0.7034	0.2349	
VELM_S	0.1189	0.923	**0.0474**

B. *Performance Metric and its evaluation.*

The results of binary classification can be categorized as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The overall accuracy is calculated by the following formula.

$$Overall\ Accuracy = \frac{\#\,TP + \#TN}{\#\,Samples}$$

$$Gmean = \sqrt{\frac{\#TP}{\#TP + \#FN} * \frac{\#TN}{\#TN + \#FP}}$$

Here, # represents number of.

C. *Parameter Settings*

Sigmoid activation function used for hidden layer neurons for both ELM and its variant. For comparison of simple majority and unweighted soft voting same set of 50 classifiers are used. This is done to have a fair comparison between them. Results presented in this section are averaged over 10 trials. In each trail 50 ELM classifiers are generated and the outputs of these classifiers are provided to VELM and UCSSV-ELM. The ensemble using unweighted class specific soft voting with ELM and ELM variant using sigmoid function are respectively represented by UCSSV-ELM and UCSSV-ELM_S. The ensemble using ELM as base classifier and simple majority voting is VELM whereas, the ensemble using ELM variant with sigmoid function to obtain probabilistic output is termed as VELM_S. Optimal number of NHN for VELM has been found by varying NHN from [10, 20… 100]. The final outcome of VELM is the majority voting of all these 50 classifiers and the final outcome of UCSSV-ELM is computed as per above algorithm.

D. *Experimental Results*

Average Overall Accuracy, AOA and mean of Gmean of the 4 classifier ensembles for various datasets are shown in Table III and Table IV respectively. It can be observed from Table III unweighted soft voting gives better results than simple majority voting. For further comparison of proposed classifiers with V-ELM, wilcoxon signed rank test is conducted. The threshold value of alpha is taken as 0.1. The p values obtained by the test are shown in Table V. The smaller p-value indicates significant improvement.

V. Conclusion and Future Work

This paper proposes a new classifier, UCSSV-ELM which is an extension of VELM. UCSSV-ELM uses unweighted (equally weighted) class specific soft voting. This work compares simple majority voting and unweighted class specific soft voting for aggregating the output of component classifiers. UCSSV-ELM performs better than VELM as can be seen from the results of wilcoxon signed rank test. Also, UCSSV-ELM_S is better than VELM_S. In general we can see that unweighted class specific voting performs better than simple majority voting. This work also studies whether the output of ELM should be converted in probabilistic output by using sigmoid activation function or the output of ELM can be directly used as input for class specific soft voting. It can be seen from the results of wilcoxon test that UCSSV-ELM is equivalent to UCSSV-ELM_S. Also VELM is equivalent to VELM_S. From this we can conclude that is is not necessary to convert the output of ELM by using probabilistic sigmoid function

VI. References

[1] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70. pp. 489–501, 2006.

[2] L. Breiman, "Bagging predictors: Technical Report No. 421," 1994.

[3] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," *Mach. Learn. Work. …*, 1996.

[4] J. Cao, Z. Lin, G. Bin Huang, and N. Liu, "Voting based extreme learning machine," *Inf. Sci. (Ny).*, vol. 185, pp. 66–77, 2012.

[5] J. Cao, S. Kwong, R. Wang, X. Li, K. Li, and X. Kong, "Class-specific soft voting based multiple extreme learning machines ensemble," *Neurocomputing*, vol. 149, Part , no. 0, pp. 275–284, Feb. 2015.

[6] N. Liu and H. Wang, "Ensemble based extreme learning machine," *IEEE Signal Process. Lett.*, vol. 17, pp. 754–757, 2010.

[7] J. Zhai, H. Xu, and X. Wang, "Dynamic ensemble extreme learning machine based on sample entropy," *Soft Computing*, vol. 16. pp. 1493–1502, 2012.

[8] Y. Lan, Y. C. Soh, and G. Bin Huang, "Ensemble of online sequential extreme learning machine," *Neurocomputing*, vol. 72. pp. 3391–3395, 2009.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[9] G. Wang and P. Li, "Dynamic Adaboost ensemble extreme learning machine," in *ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings*, 2010, vol. 3.

[10] L. Yu, X. Xiujuan, and W. Chunyu, "Simple ensemble of extreme learning machine," in *Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP'09*, 2009.

[11] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, pp. 255–287, 2011.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka

Zainal, K.
Faculty of Science & Technology
Islamic Science University of Malaysia (USIM)
Nilai, Negeri Sembilan, Malaysia

Sulaiman, N.F.
Faculty of Science & Technology
Islamic Science University of Malaysia (USIM)
Nilai, Negeri Sembilan, Malaysia

Jali, M.Z.
Faculty of Science & Technology
Islamic Science University of Malaysia (USIM)
Nilai, Negeri Sembilan, Malaysia

***Abstract*—This paper reported and summarized findings of spam management for Short Message Service (SMS) which consists of classification and clustering of spam using two different tools, namely RapidMiner and Weka. By using the same dataset, which is downloaded from UCI, Machine Learning Repository, various algorithms used in classification and clustering in this simulation has been analysed comparatively. From the simulation, both tools giving the similar results that the same classifiers are the best for SMS spam classification and clustering which are outperformed than other algorithms.**

Keywords-SMS spam; RapidMiner; Weka; Naïve Bayesian (NB); Support Vector Machine (SVM); k-Nearest Neighbour (kNN); K-Mean; Cobweb; Hierarchical clustering; spam classification; spam clustering.

I. INTRODUCTION

Issue of spam has been widely discussed all over the world. Impact caused by spam has been noticing as extremely risky. Nowadays, spam does not only apply to email form but quite numerous to mention such as web spam, SMS spam and instant messaging spam. Spam has been evolved along as the technology advances.

This paper focuses only on SMS text spam, which covers the issue and available technology for spam management. Data mining is the strategy employed as a part of this paper, whereby it is use of automated data analysis techniques to reveal previously undetected relationships among data items. This regularly includes the analysis of data stored in a library or data warehouse. Three of the major data mining techniques are regression, classification and clustering. The heart of spam management basically includes two phases of data mining; classification and clustering of spam.

This paper basically is arranged in sections, as follows. Section 2 summarizes related works in this field; Section 3 explains the methodology applied in this simulation. Section 4 elaborates tools used and the description of the dataset is explained in Section 5. The design of this experiment is justified in Section 6 and the finding of experiment is discussed in the subsequent section together with the conclusion of the simulation.

This research is funded by RAGS, under Ministry of Education, Malaysia.

II. LITERATURE SURVEY OF RELATED WORKS

Technology has been progressively widened in many aspects; from internet to mobile. This technological advance has affect and influences to the people's lifestyle globally in many ways including communication channels, news release, shopping pattern and much more aspects of daily life. The most outstanding technology now is mobile phone; as the usage of mobile phone has been highly propagate these recent years. Phone calls, text messaging or SMS and accessing Internet are the most common functions of a mobile phone. But almost a couple of decades ago and still, the use of SMS have been definite nuisance to users all over the world because of the misuse of SMS by unscrupulous parties, which is called as spam.

Spam management consists of at least three main phases; classification, clustering and severity determination level, as suggested in Sulaiman et al. [1]. Classification of spam messages as the main process is significantly able in assisting users to differentiate messages between of ham and spam. Identification of spam messages would reduce the possibility of risk since users will simply ignore the message. While as for spam clustering, it is important to find the current trend of spam dissemination. This information is kindly useful to project the possible amount of loss such as phishing contents might have a higher impact of risk comparing to message with free ringtone offer.

Many researches have been done and it is still an ongoing process in developing filtering spam. Mahmoud et al. [2] has developed a framework to filter SMS spam using a novel introduced algorithm that is mimic human body defence systems knowingly as Artificial Immune System (AIS). In their analysis, findings showed that their proposed engine is giving 91% of accuracy rate compared the performance of Naïve Bayesian (NB) with only 88%.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Rafique et al. [3] proved that Support Vector Machine (SVM) is giving 93%, the highest accurate value in their developed spam filtering framework, compared with other algorithms such as NB, C4.5 and Repeated Incremental Pruning to Produce Error Production (RIPPER).

Cai et al. [4] had developed a system for spam detection using Winnow algorithm. This experiment particularly executed with Chinese language SMS messages. Although the experimental results illustrated that this system works well, it is possible that the system could be further enhanced by developing the feature selection method and the decision making procedure.

III. Methodology: Supervised And Unsupervised Machine Learning Algorithms

Classification of spam is referring to process of messages differentiation between spam and ham (valid or non-spam messages), while for clustering, it is a process of partitioning messages into cluster or group according to its similarity features or characteristics [5].

Both processes can be accomplished with the aid of machine learning with the establishment of various algorithms. Machine learning algorithms are structured into taxonomy that is based on the desired outcome. It is a subfield from the broad field of artificial intelligence, which intend to make machines be able to learn like human [6]. Common algorithm used include supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, transduction and learning to learn machine learning [7].

With reference to Donalek [5], classification process is implemented using supervised machine learning, while clustering applying unsupervised machine learning. As this experiment involved a process of classification and clustering spam messages, hence both supervised and unsupervised machine learning are applied.

Algorithms that are chosen to be applied in tools of RapidMiner and Weka are elaborated in the following paragraphs. All of these six chosen algorithms are prominent for its performance in data mining field and available in both tools.

A. Classification Implemented Using Supervised Machine Learning Algorithms

According to Brownlee [8], supervised learning deploy an input data that is called training data and has a pre-defined label or result, for example spam / not spam or a stock price at a time. A model is arranged through a training process where it is obliged to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Example problems that apply supervised learning are classification and regression. The following highlights on various supervised learning algorithm used in this simulation.

1) Naïve Bayesian

Naïve Bayesian (NB) classifier is a simple probabilistic classifier based on concerning Bayes's hypothesis with strong independence theory. A more descriptive term for the underlying probability model would be an 'independent feature model'. The NB inducer computes conditional probabilities of the classes given the instance and picks the class with the highest posterior. Depending on the precise nature of the probability model, NB classifier can be trained very efficiently in a supervised learning setting [9].

According to Awad et al. [6], NB classifier was proposed for spam recognition in 1998. In NB, word probabilities play the main rule which every word has certain possibility of occurring in spam or ham messages in its database. If some words occur often in spam but not in ham, then an incoming message is probably a spam.

2) Support Vector Machine

Support Vector Machine (SVM) is a learning algorithm with 2-class classification method. This algorithm converts miscellaneous domain knowledge with overlapping inputs into non-overlapping parametric objects by modelling the instances from the input space to the feature space using kernel functions. The classification is done by constructing a hyper plane between instances of different classes [3].

According to El-halees [10], y which classifies messages as spam or legitimate according to following dot product:

$$y = w.x - b \tag{1}$$

where x is a feature vector of messages composed of words. w is the weight of corresponding x. b is a bias parameter determined by training process.

3) k-Nearest Neighbour

k-Nearest Neighbour (kNN) classifier is considered as an example-based classifier, means the training documents are used for comparison, rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document need to be categorized, the k most similar documents (neighbours) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not. Additionally, finding the nearest neighbours can be accelerated using traditional indexing methods. To decide whether a message is spam or ham, it is referring to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the class of the messages that are closest to it. The comparison between the vectors is a real time process. This is the idea of kNN algorithm [6].

B. Clustering Implemented Using Unsupervised Machine Learning Algorithms

Brownlee [8] defined unsupervised learning has an input data that is not labelled and does not have a known result. A model is prepared by deducing structures present in the input data and example problems are association rule learning and clustering. The following highlights on various unsupervised machine learning algorithm used in this simulation.

1) K-Means

K-Means clustering algorithm was first proposed by Macqueen in 1967 which was uncomplicated, non-supervised learning clustering algorithm. K-means is a partitioning clustering algorithm, this technique is used to classify the given data objects into *k* different clusters through iterative method, which tends to converge to a local minimum. So, the outcomes of generated clusters are dense and independent of each other [11].

The K-Means algorithm is the best known squared error-based clustering algorithm [12]. It involves the processes of:

- Selection of the initial *k* means for *k* clusters;
- Calculation of the dissimilarity between an object and the mean of a cluster;
- Allocation of an object to the cluster whose mean is nearest to the object; and
- Re-calculation of the mean of a cluster from the objects allocated to it so that the intra cluster dissimilarity is minimized.

2) Cobweb

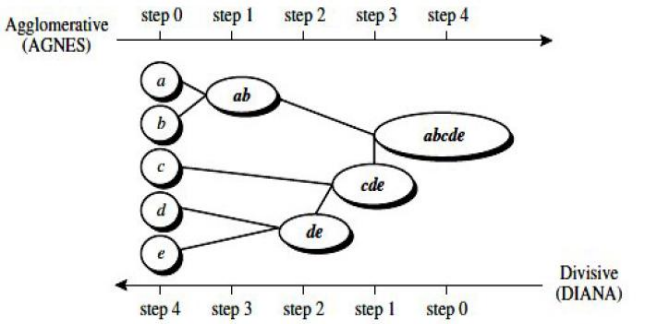
The Cobweb algorithm was developed by machine learning researchers in 1980s for clustering objects in an object-attribute dataset. This algorithm yields a clustering dendrogram called classification tree that differentiates each cluster with a probabilistic description. Cobweb generates hierarchical clustering where clusters are described probabilistically [13].

3) Hierarchical

Hierarchical method creates a hierarchical decomposition of the given set of data objects forming a dendrogram - a tree which splits the database recursively into smaller subsets. The dendrogram can be formed in two ways, bottom up or top down. Hierarchical algorithm combines or divides existing groups, creating a hierarchical structure that reflects the order in which groups are merged or divided.

The bottom up approach, also called the agglomerative approach, starts with each object forming a separate group. It successively merges the objects or groups according to some measures like the distance between two centres of two groups and this is done until all of the groups are merged into one, or until a termination condition holds.

The top down also called the divisive approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters accordingly to some measures until eventually each object is in one cluster, or until a termination condition holds.



Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

Figure 1. Hierarchical clustering process [14].

IV. TOOLS

These whole activities of spam classification and clustering can be done with the aid of established algorithms that has been developed using software. These tools are available as commercial or free products.

As for the implementation of this experiment, two tools with free license (freeware) has been used in this simulation, RapidMiner (Community Edition) and Weka. Both tools are prominent as data mining software and can be downloaded from the Internet.

Christa et al. [15] in their paper evaluated and analyzed comparatively of various data mining tools such as KNIME, RapidMiner, Weka, Tanagra and Orange. These tools has its own advantages and unique. Among all these data mining tools, Weka and RapidMiner have the biggest and most active user communities. Both of them quickly implement (and integrate) new and emerging machine learning algorithms into their systems.

As in this paper, the simulation process is using RapidMiner and Weka as the data mining tools.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. RapidMiner

RapidMiner is data mining software, which can be use as a standalone application for data analysis or integrate as a data-mining engine into other products. This tool has unique features such as:

- Data integration, analytical Extract Transform Load (ETL), data analysis and reporting into a single suite;
- Powerful intuitive Graphical User Interface (GUI) for the design of analytical processes;
- Repository for process, data and metadata management;
- Metadata transformation which results inspection available during design;
- Support on-the-fly error detection and quick fixes; and
- Complete and flexible with hundreds of methods available for data integration, data transformation, modelling and visualization.

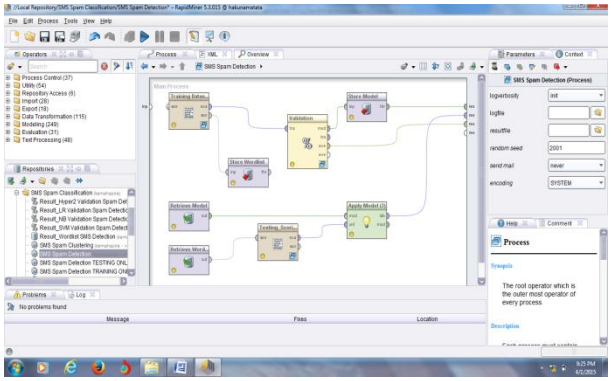


Figure 2. Interface of RapidMiner.

B. Weka

Weka is a collection of machine learning algorithms for data mining tasks with GUI. This application is named after a flightless bird of New Zealand that is very inquisitive. The algorithms can either be applied directly to a dataset or called from own Java code. Weka contains feature for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes. There are four buttons of GUI in Weka [16] which are:

- Explorer – an environment for exploring data
- Experimenter – an environment for performing experiments and conducting statistical tests between learning schemes
- Knowledge Flow – this environment supports essentially the same functions as the Explorer but with a drag and drop interface and it supports incremental learning
- Simple CLI – provides a simple command line interface that allows direct execution of Weka commands.

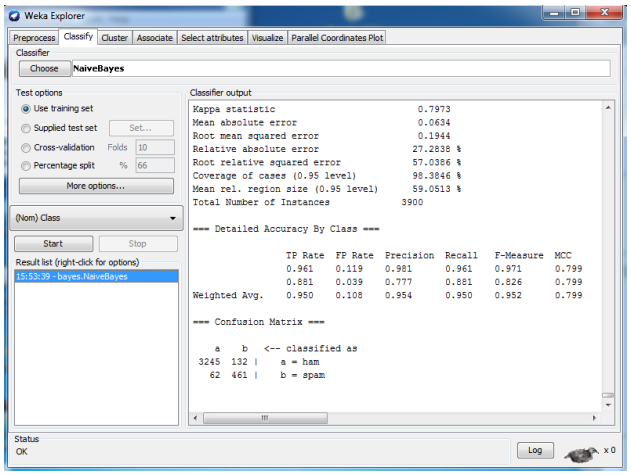


Figure 3. Interface of Weka.

Both of the aforementioned tools are deployed in this simulation by an application of specified algorithms that has been stated earlier.

As to summarize this experiment, this simulation is depicted in the following figure that exhibit the integration between spam management processes, machine learning algorithms and data mining tools; RapidMiner and Weka.

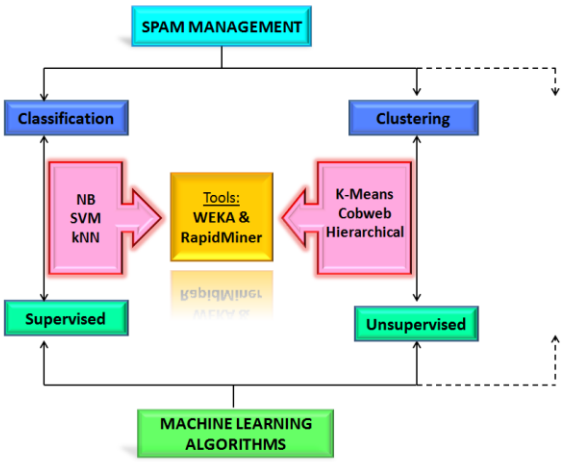


Figure 4. Integration of spam management processes with machine learning algorithms and experimental tools.

V. DATASET DESCRIPTION

This experiment is using a dataset downloaded from UCI, Machine Learning Repository. This corpus site stored a collection of public set of SMS messages labelled as spam and ham (non-spam) for the use by many researchers. The dataset downloaded consists of 5,572 instances which is 4,825 messages are labelled as ham and 747 as spam.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

All 5,572 downloaded SMS messages from this corpus are used both for training and testing phases with the fraction of 7:3 (7 for training and 3 for testing). 70% of both ham and spam messages are used in training phase, as the more the dataset use for training, the better the model would be when it is applied in the testing phase. The other 30% of both ham and spam messages are used in testing phase. This dataset of SMS messages are divided as follows:

TABLE I. ALLOCATION OF THE DATASET INTO TWO PARTS, TRAINING AND TESTING PROCESS

	Training phase (70%)	Testing phase (30%)	Total messages
Labelled as HAM	3,377	1,448	4,825
Labelled as SPAM	523	224	747
Total	**3,900**	**1,672**	**5,572**

As to provide a fair comparison, this same dataset is deployed using six aforementioned algorithms in RapidMiner and Weka.

VI. EXPERIMENTAL DESIGN

As stated earlier, spam management involved three main processes, which are spam classification, clustering and severity determination of the detected spam. However, the focal objective of this experiment is focusing on the first two processes; namely classification and clustering. Different classifiers or algorithms models are applied in RapidMiner and Weka (only used Explorer as GUI application), purposely to find the best suited algorithm for those two processes.

While as for the data mining tools, this experiment is using RapidMiner version 5.3.015 and Weka version 3.7.10.

A. Spam Classification

As for the classification process, there are two levels that applied: training and testing level. These two different levels are actually reflected the supervised machine learning characteristics.

During training, a set of 3,377 ham and 523 spam labelled messages are running through a few different classifiers separately. Then, for testing, based on stored correlation attributes during the training level, a set of unlabelled messages are running through those classifiers. Finally the results of these findings are verified to measure its performance.

In spam classification process, there are four methodologies that the testing of unlabelled messages has been executed, as explain in Table II. All classifiers had been re-run in Method 1 and 2, while Method 3 and 4 only re-run using the best classifier identified in Method 1 and 2, since the result of best classifier in Method 3 and 4 remain the same as in Method 1 and 2.

TABLE II. DESCRIPTION OF EXECUTION PLAN FOR TRAINING AND TESTING PHASE

Method	No. of messages		Description of testing phase
	Training phase	**Testing phase**	
1	3,900 messages labelled with ham or spam	1,672 unlabelled messages	These 2 phases of spam detection are run simultaneously. The time taken to complete the whole process is recorded.
2	3,900 messages labelled with ham or spam	1,672 unlabelled messages	These 2 phases of spam detection are run separately. The time taken to complete these 2 different processes is recorded.
3	3,900 messages labelled with ham or spam	1 unlabelled message	These 2 phases of spam detection are run separately. As for the testing phase, a few different unlabelled messages are run repeatedly, with only 1 unlabelled message run at one time (reflect the actual environment).
4	100 messages labelled with ham and spam 1,250 messages labelled with ham and spam 2,500 messages labelled with ham and spam	1,672 unlabelled messages	These 2 phases of spam detection are run separately. As for the testing phase, 1,672 unlabelled messages are run repeatedly with a different size of database stored during training phase. This method is to demonstrate the link between spam library / database (developed during training) with unlabelled messages run in testing phase. Also to find the influence degree of number of messages used in training with the result of spam classification, in term of accuracy rate.

B. Spam Clustering

Clustering is concerned with grouping together SMS spam messages that are similar to each other and dissimilar to the other clusters or groups. Clustering is a technique for extracting information from unlabelled data. This is important as to learn the pattern of spam content.

At this phase, all SMS messages that have been pre-classified as spam are used for further process, to be cluster according to its category by referring to the content of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

messages. Since clustering process is employing an unsupervised machine learning algorithm, there is not a requirement to run a training dataset. Therefore, all 747 spam messages from the dataset are used to be further tested with different classifiers. Referring to Delany et al. [17], these 747 spam messages have been categorized into 10 clearly defined groups, which are:

- Competitions
- Chat
- Claims
- Dating
- Prizes
- Services
- Finance
- Ringtones
- Voicemail
- Miscellaneous

VII. Performance Measurement

In order to rank or select the best classifier in SMS spam classification, some of the performance measurements need to be choosing to determine the best option.

As to measure the best performance, the higher the value of Accuracy, the better the classifier it is. The time taken to complete the process also selected as one of the criteria in performance measurement. The shorter the time taken, the better the classifier it is. The following measurements are used to formulate the performance that commonly used in machine learning [18].

- Accuracy in percent (%): the value of spam being correctly classified. It is also reflect the overall performance of the framework.

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

- True Positive (TP): the number of SMS spam classified as spam.
- True Negative (TN): the number of SMS ham classified as ham.
- False Positive (FP): the number of SMS ham falsely classified as spam.
- False Negative (FN): the number of SMS spam falsely classified as ham.

Processing time in seconds: the time taken to complete the process. The shorter the time taken the better the classifier it is.

TABLE III. Confusion Matrix Of Classification Algorithm

		Prediction	
		Spam	Ham
True	Spam	**TP**	**FN**
	Ham	**FP**	**TN**

VIII. Results and Discussion

A. Spam Classification

These 1,672 instances of dataset have constructed 5,209 regular attributes in RapidMiner. The result of SMS spam classification for unlabelled 1,672 messages (1,448 ham and 224 spam messages) that running through with different classifiers in both tools are table out as below:

Method 1:

TABLE IV. Accuracy Rate and Time Taken To Process Messages (Training And Testing Phases Executed Simultaneously)

	Accuracy (%)		Processing time (seconds)	
	RapidMiner	**Weka**	**RapidMiner**	**Weka**
NB	84.79	94.56	279	0.91
SVM	96.64	98.21	263	2.48
kNN	94.74	94.80	802	0.01

Method 2:

TABLE V. Accuracy Rate And Time Taken To Process Messages (Training And Testing Phases Executed Separately)

	Accuracy (%)		Processing time (seconds)			
	Rapid Miner	**Weka**	**RapidMiner**		**Weka**	
			Training	**Testing**	**Training**	**Testing ******
NB	84.79	95.03	173	38	0.64	-
SVM	96.64	99.33*	195	21	1.54	-
kNN	94.74	99.85	1091	493	6.0	-

*Even though kNN is giving a slightly higher accuracy rate compared to SVM in Weka, time taken to process messages is still significantly lesser than SVM.
**means there is no record of time taken for testing phase in Weka.

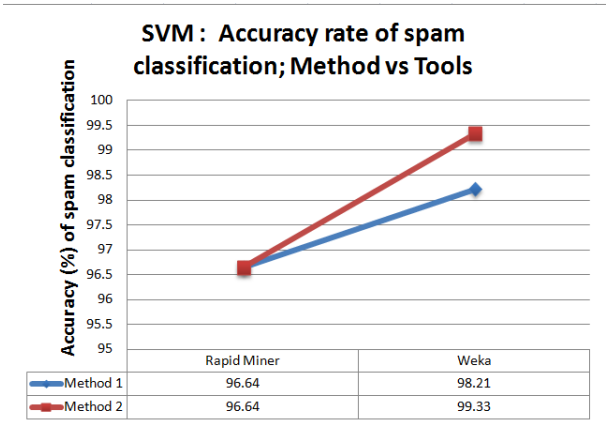


Figure 5. Accuracy rate of spam classification for Method 1 and Method 2 using RapidMiner and Weka.

As for spam classification, it showed that SVM is the best classifier in both RapidMiner and Weka.With referring

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

to graphical representation in Figure 5, RapidMiner giving the same accuracy rate (96.64%) of spam classification in both methods (training and testing phases executed simultaneously and separately), but different in processing time. On the other hand, deployment of Method 2 in Weka is giving a slightly higher of accuracy rate (99.33%) in spam classification compared to Method 1 (98.21%).

Method 3:

TABLE VI. COMPARISON OF PREDICTIVE RESULTS AND TRUE LABEL OF UNLABELLED MESSAGES USING SVM CLASSIFIER IN RAPIDMINER

	Processing time (seconds)		Description of findings in testing phase		
	Training phase	Testing phase	Unlabelled message	Prediction by classifier	True label
SVM	195	0	202.txt	Ham	Ham
			567.txt	Ham	Ham
			862.txt	Ham	Ham
			1198.txt	Ham	Ham
			1443.txt	Ham	Ham
			1578.txt	Spam	Spam
			1599.txt	Spam	Spam
			1616.txt	Spam	Spam
			1658.txt	Spam	Spam
			1670.txt	Spam	Spam

TABLE VII. COMPARISON OF PREDICTIVE RESULTS AND TRUE LABEL OF UNLABELLED MESSAGES USING SVM CLASSIFIER IN WEKA

	Processing time (seconds)		Description of findings in testing phase		
	Training phase	Testing phase	Unlabelled message	Prediction by classifier	True label
SVM	1.51	-	One.arff	Ham	Ham
			Two.arff	Ham	Ham
			Three.arff	Ham	Ham
			Four.arff	Ham	Ham
			Five.arff	Ham	Ham
			Six.arff	Spam	Spam
			Seven.arff	Spam	Spam
			Eight.arff	Spam	Spam
			Nine.arff	Spam	Spam
			Ten.arff	Spam	Spam

Messages are in separate text file and randomly choose to be further tested. Results show 100% accurate, whereby prediction by classifier are match with the actual label of the messages.

Method 4:

TABLE VIII. RESULTS OF TESTING PHASE (ACCURACY AND TIME TAKEN) WHEN VARIOUS DATABASE ARE DEPLOYED USING SVM IN RAPIDMINER

	No. of messages used in training phase	Processing time (seconds)		Accuracy (%)
		Training	Testing	
SVM	100	15	7	65
	1,250	69	14	93.28
	2,500	138	16	94.72
	3,900	195	21	96.64

Referring to Table I, as for the unlabelled messages in testing phase, 1,448 messages is ham and 224 messages is spam.

TABLE IX. RESULT OF TESTING PHASE (ACCURACY AND TIME TAKEN) WHEN VARIOUS DATABASE SIZE ARE DEPLOYED USING SVM IN WEKA

	No of messages used in training phase	Processing time (seconds)		Accuracy (%)
		Training	Testing	
SVM	100	0.02	-	99.0
	1,250	0.25	-	94.96
	2,500	0.48	-	96.08
	3,900	1.54	-	99.33

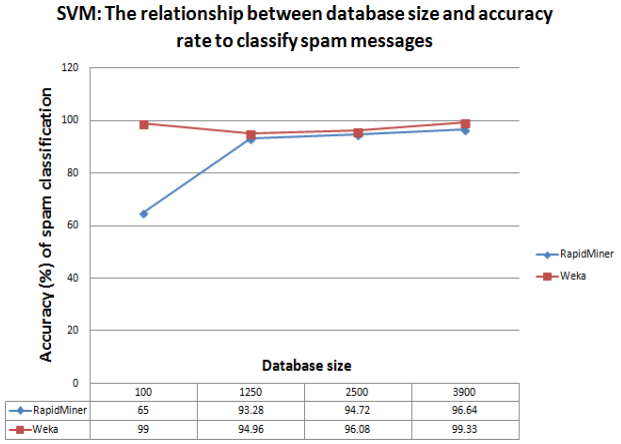


Figure 6. The relationship between database size and accuracy rate of spam classification using RapidMiner and Weka (Method 4).

As referring to Figure 6, this experiment showed that the more the labelled messages are deployed during the training phase as to establish the database library, the better the classifier would become. This result also showed again that Weka is giving a slightly higher accuracy rate compared to RapidMiner.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Spam Clustering

As explained in paragraph 6.2, there are 10 groups of spam messages that have been pre-defined of its category. Hence, as for the number of cluster to be defined in every classifier is chosen as 10.

TABLE X. CLUSTERING OF 747 SPAM MESSAGES USING 3 DIFFERENT ALGORITHMS IN RAPIDMINER AND WEKA

	Results of prediction		Processing time (seconds)	
	RapidMiner	**Weka**	**RapidMiner**	**Weka**
K-Means	C0: 343 items C1: 47 items C2: 83 items C3: 17 items C4: 50 items C5: 20 items C6: 39 items C7: 33 items C8: 66 items C9: 49 items	C0: 212 items C1: 2 item C2: 15 items C3: 2 items C4: 6 items C5: 2 items C6: 501 items C7: 2 items C8: 2 items C9: 3 items	37.0	2.7
Hierarchical	Unable to cluster messages.	C0: 729 items C1: 2 items C2: 2 items C3: 2 items C4: 2 items C5: 2 items C6: 2 items C7: 2 items C8: 2 items C9: 2 items	-	0.94
Cobweb	C0:747 items	C0:747 items	31.0	11.22

*C = Cluster

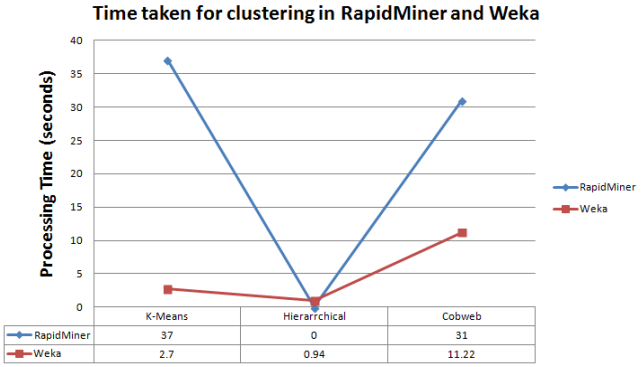


Figure 7. Time taken (seconds) for clustering using K-Means, Hierarchical and Cobweb; in RapidMiner and Weka.

The findings suggest that, with the use of RapidMiner and Weka in SMS spam classification and clustering, it showed that:

- SVM classifier is the best to be used in SMS spam classification, which the result of testing for 1,672 unlabelled messages produced in 21 seconds using RapidMiner with 96.64% of accuracy and 1.54 seconds with 99.33% of accuracy using Weka; and
- K-Means algorithm is the best suited to cluster 747 spam messages into 10 groups in 37 seconds using RapidMiner and 2.7 seconds using Weka.

IX. CONCLUSION

The summary of the comparison in spam classification and clustering using RapidMiner and Weka are tabulated in Table XI.

TABLE XI. THE COMPARISON BETWEEN RAPIDMINER AND WEKA

Phase of Spam Management	**WEKA**		**RapidMiner**	
	Best classifier	**Accuracy & Time taken**	**Best classifier**	**Accuracy & Time taken**
Spam Classification	SVM	Accuracy: 99.33% Time taken: 1.54 seconds.	SVM	Accuracy: 96.64% Time taken: 21 seconds for testing phase only for separate execution.
Spam Clustering	K-Means	Time taken: 2.7 seconds.	K-Means	Time taken: 37.0 seconds.

According to the XI, this experiment demonstrated that SVM is the best classifier for spam classification and K-Means is the most suitable algorithms to cluster spam messages. These algorithms giving a promising result both in RapidMiner and Weka. Other than that, this experiment also shows:

- As elaborated in Method 2 findings, it suggest that the testing phase will take shorter time when it is run separately with training phase;
- As the size of datasets increases, time taken to classify the messages are also increase;
- Logically, in practice only one or two SMS will be delivered to mobile phone at one time, and this will not consume much time in classification process to detect either it is a spam or not, as suggest in the findings of Method 3;
- As suggest in Method 4, the more the dataset used in training phase, the better the classifier will perform and giving a higher accuracy rate but as the size of datasets increases, the time taken to 'learn' and classify the spam messages also increases in both training and testing phases;
- Weka tool is giving the shortest time in executing the spam classification and clustering and also giving a higher rate of accuracy which is significantly better compared to RapidMiner;
- Both tools resulted the same sequence of highest accuracy in spam classification; SVM, KNN and NB;
- Cobweb is not a suitable algorithm to cluster SMS messages. All 747 spam messages were clustered into one group only instead of 10 groups, both in RapidMiner and Weka; and
- This study revealed that the same classifier performed dissimilarly when run on the same dataset but using different tools.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The main motivation for different classification algorithms is resulting in high accuracy rate. Each method has its own variety of algorithms. Various algorithms of these methods were used to predict the pattern and behaviour of the dataset, as in this case is spam messages.

This similar simulation can be executed for other different and advanced algorithm. An employment of Artificial Immune System (AIS) is one of the techniques that can be further considered since this AIS has been well developed and matured to be tested in many field to detect malicious behavior such as email spam classification, virus and intrusion detection.

ACKNOWLEDGMENT

Authors wish to thank the Ministry of Education, Malaysia for funding this research. The fund is known as RAGS, under the code number RAGS/2013/USIM/ICT04/2.

REFERENCES

[1] Sulaiman, N. F., and Jali, M. Z. "Integrated Mobile Spam Model Using Artificial Immune System Algorithms", Knowledge Management International Conference (KMICe), 2014.

[2] Mahmoud, T. M., and Mahfouz, A. M. "SMS Spam Filtering Technique Based on Artificial Immune System", International Journal of Computer Science, 2012.

[3] Rafique, M. Z., Alrayes, N., and Khan, M. K. 2011. Application of Evolutionary Algorithms in Detecting SMS Spam at Access Layer.

[4] Cai, J., Tang, Y., and Hu, R. "Spam Filter for Short Messages using Winnow", International Conference on Advanced Languange Processing and Web Information Technology, 2008.

[5] Donalek, C. 2011. Supervised and Unsupervised Learning.

[6] Awad, W. A., and ELseuofi, S. M. "Machine Learning Methods for Spam Email Classification", International Journal of Computer Science and Information Technology, 2011.

[7] Ayodele, T. O. 2010. Types of Machine Learning Algorithms. In New Advances in Machine Learning.

[8] Brownlee, J. 2013. A Tour of Machine Learning Algorithms. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

[9] Lakshmi, R. D., and Radha, N. "Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools", International Journal on Computer Science and Engineering, 2010.

[10] El-halees, A. "Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques", International Arab Journal of Information Technology, 2009.

[11] Sehgal, G., and Garg, D. K. "Comparison of Various Clustering Algorithms", International Journal of Computer Science and Information Technologies, 2014.

[12] Saxena, P., and Lehri, S. "Analysis of Various Clustering Algorithms of Data Mining on Health Informatics", International Journal of Computer and Communication Technology, 2013.

[13] Sharma, N., Bajpai, A., and Litoriya, R. " Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, 2012.

[14] Godara, S. "A Comparative Performance Analysis of Clustering Algorithms", International Journal of Engineering Research and Applications.

[15] Christa, S., Madhuri, K., and Suma, V. 2012. A Comparative Analysis of Data Mining Tools in Agent Based Systems.

[16] Chaudhari, B., and Parikh, M. " A Comparative Study of Clustering Algorithms Using Weka tools", International Journal of Application or Innovation in Engineering and Management (IJAIEM), 2012.

[17] Delany, S. J., Buckley, M., and Greene, D. 2012. SMS Spam Filtering: Methods and Data. Expert Systems with Applications , Elsevier, 01(10).

[18] Wang, A. H. 2012. Machine Learning for the Detection of Spam in Twitter Networks. *ICETE*, 319–333.

AUTHORS PROFILE

Kamahazira Zainal currently is pursuing her Ph.D in Science and Technology specifically in Information Security and Assurance from Islamic Science University of Malaysia (USIM). Previously received her Master in Information Security from Universiti Teknologi Malaysia (UTM) in 2008 and Bachelor of Computer and Communication Engineering from Universiti Putra Malaysia (UPM) in year of 2000.

Nurul Fadhilah Sulaiman presently studying Master in Information Security from Islamic Science University of Malaysia (USIM). She received her degree in Computer Science (Information Security and Assurance) in 2013, also from USIM. She already published a paper related to this field titled "Integrated Mobile Spam Model Using Artificial Immune System Algorithms" in Knowledge Management International Conference 2014 (KMICe 2014).

Dr. Mohd Zalisham Jali received his Ph.D in 2011 from the Plymouth University, UK. Dr Zalisham is now senior lecturer of the Computer Science program. His current research of interest includes information security, web accessibility and e-learning.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Security Architecture with NAC Using Crescent University as Case Study

Ayangbekun Oluwafemi J.
Department of Information Systems
University of Cape Town
Cape Town, South Africa

Audu Zainab O.
Department of Computer Science
Crescent University Abeokuta
Ogun State, Nigeria.

***Abstract*— A campus network implemented for a university is a medium for effective communication in such environment. The major advantage of having such network is the shared nature of resources and mobile nature of students, teachers and administrators. This research based on the low level of network capacity in Crescent University taking into consideration applying NAC and some selected protocols over both wired and wireless network. Star topology was employed to design the network, a network with centralized server because of the size of the school and the limitations it might encounter if other forms of technologies are deployed, as illustrated in the diagrams. Programmable switches was used to illustrate how the virtual devices and endpoint devices can be used to handle security threats. The use of Wired Equivalent Privacy (WEP) or WPA (Wi-fi Protected Access) which is a security algorithm for IEEE 802.11 wireless networks and other protocols for wired devices to target security challenges. As new ways to improve and achieve goals in a university is increasing, using modern technologies has made learning and teaching easier in which few universities have successfully adopted a standard networking model across Africa.**

Keywords- Network Access Control (NAC), Switch, campus network, protocols

I. INTRODUCTION

Campus networks harbour many devices that can be connected through a wired or wireless means which makes it exposed to threats that can compromise the network. It is quite difficult to monitor the activities of end users on a network especially in developing countries of the world where the case study is located. However, the use of network devices to restrict and monitor users by giving out policies or administrative configurations can help keep the network safe and provide effective IT support.

The institution needs to store large amount of data and also share some of these resources, manipulate them and protect them from external threats to the network. Increased number of users leads to more devices: laptops, tablets and mobile phones which calls for more security attention. This would not be a worry if a form of policy is set down such as authentication, security compliance, control of end users, authorized access. Students are very inquisitive as to what they can try out and sometimes can lead to a security breach. The sensitive nature of this organization requires a system that supports reliability, increased bandwidth, redundancy, network convergence and flexibility. This can be achieved by customizing protocols to fit the existing system of information and communication technology in the university. For such to be implemented, Cisco switches act as the backbone of this solution to effectively share resources within the network. Rather than relying solely on end devices on security solution which can be easily compromised, the network should provide an administrative control for higher level of security and maintenance.

II. LITERATURE REVIEW

Network attacks have been discovered to be as varied as the system they attempt to penetrate. Attacks are known to either be intentional or unintentional and technically competent intruders have been interested in targeting the protocols used for secure communication between networking devices [1]. This review addresses the suitable approach to create a network in a university and to curb threats to such networks. `
The level of success of implementation of modern networking is quite low in developing countries as compared to developed countries. The bridge between both sides in terms of learning and education however, is increasing by the day.
The term “campus” is a building or group of buildings all connected into one enterprise network that consists of many local area networks (LAN) it is constrained to a fixed geographic area.
The need for new forms of technology is important but it comes with a great deal of risks which solutions have to be provided to show improvements. Relative to cost, low bandwidth, more competent IT staffs; the major problem of providing a model is security which is addressed in this research.
Apart from the ability of students to be sound in ICT starting from their foundation, it makes it easier to navigate outside the school walls in organization settings. Most companies would not settle for less when the employees have a solid knowledge

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

of a networked environment. This is the case for a country like Nigeria and other West African countries.

There are various tools that are employed to provide security to networks such as antivirus, SSH protocol, port-security, VPN for remote access, Network Audit, encryption among others depending on how it is carried out. The most commonly used meaning for information security is the preservation of [2];

- Confidentiality or protection from unauthorized use or disclosure of information
- Integrity, ensuring data accuracy and completeness through protection from unauthorized access, unanticipated, or unintentional modification, and including authenticity and
- Availability, making data available to the authorized users on a timely basis and when needed.

Fostering an open environment doesn't always translate to letting the inmates rule the asylum. However, for several reasons, security pros in educational institutions have turned to network access control (NAC) technology to help keep their networks safe [3].

Taking the end users into account shows that they are keener about the resources they want to access rather than the risks or threats that it can cause to the network. The end-user computing evolution provided computing power at the workplace, and resulted in end-user demand for access to corporate data with little regard for the security of that data [4].

Furthermore, according to "Cisco Reviews International Education Survey Findings", reducing administrative expenses, preventing internet abuse, adding more cyber security and enabling students and faculty to work together are the top three technology issues for the Middle East and **Africa**.

III. MODEL DESIGN

The first step of the design is to create a network mode for the university. There is an existing network which is restricted to the Information and communication Technology Laboratory. Other colleges and departments (Admissions, Bursary, Vice chancellor's office, Siwes office, library), Lecture rooms are not connected over a network. Each of these sections have their individual wireless devices to connect to the internet and no form of intranet.

A logical design is then constructed using a powerful simulator; Cisco Packet Tracer to show the logical view of the suitable network design for the institution. The (fig.1) below represents how the design is being modelled with different devices such as workstations, switch, routers etc connected to each other to share resources over a safe network.

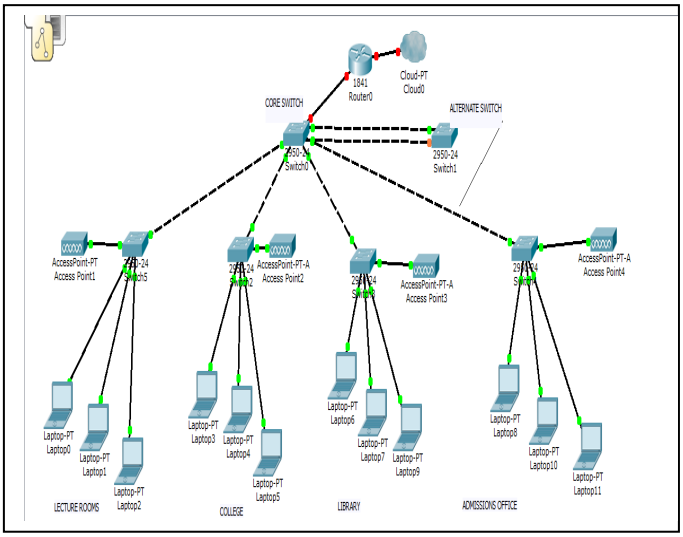


Figure 1. Institutional wired network

The above figure (fig 1) describes how the devices are connected to each other via wired means such as laptops desktops and switches. It takes lesser requirements to secure network at this stage as to the figure below (fig 2) when wireless devices are introduced to the network. As more devices gain access to the network, which in today's world is wireless, there is need for employing improved security measures.

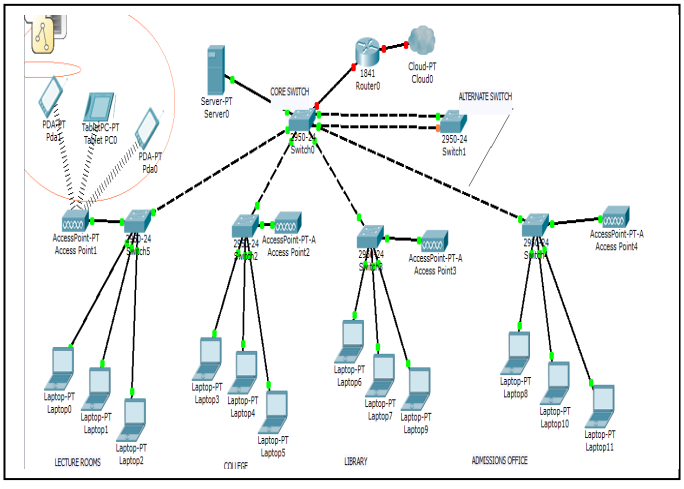


Figure 2. Wired and wireless model

It is very imperative that the average number of end users to be taken account of, number of switches and other devices required, size of the school etc should be taken into

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

consideration. However the problem of security arises from the following factors such as;

- The ease at which information in its electronic form can be accessed, manipulated and transported
- The security of users information accessed or transmitted in a networked environment

These major issues in an educational system requires a very high level of security which can be achieved by implementing various protocols in different forms and adding some security hardware and software in such networks as NAC has an advantage of interoperability with other vendors. Therefore, using cisco devices (core 6500 series, catalyst 2960-S and 3750E series) and architecture is best advised for this Boarderless network for connecting anyone, anytime, anywhere, using any device at any time. Its latest innovations of centralized policy, unified management and automated video and voice.

Thereof, control of end-user devices within an environment using NAC has a connection with all the end users and monitors their activities. As new devices are introduced to the network, it has the feature of ensuring it meets its standard on the network and regular update of the new and existing devices. These however makes the network to be security compliant and also grants both legitimate user authentication and an Authorized access to the network resources. [5]

A. MODES OF SECURING A NETWORK

There are various modes to network security over a wired/wireless domain. Some of the threats are malware, computer virus, rogue security software, Trojan horse, malicious spyware, computer worm, rootkits.

- Spanning Tree Protocol (STP) is a layer 2 protocol that runs on bridges and switches with a specification of IEEE 802.10. This protocols helps prevent loops in a redundant network.[7]
- Per VLAN Spanning Tree is a feature available on some catalyst series switches that implements a separate instance of spanning tree for each VLAN configured on a network.[8] The protocol entails having separate STP per VLAN so the flaws of a VLAN does not affect other VLANs .
- SSH (Secure Shell) is a protocol which provides a secure remote access connection to network devices. Communication between the client and server is encrypted. It is Telnet with encryption. To access the remote system will require a username and password.
- Firewall is a network security system that controls the incoming and outgoing traffic based on an applied rule set.it serves as a barrier between a trusted secure internal network and another network (internet) that could not be trusted or secure. Firewalls can be used based on the network size and other factors. Software and Hardware firewalls like Microsoft ISA server and Linksys respectively are types that support interoperability between devices. They act as DHCP server and so the existing DHCP (Dynamic Host Configuration Protocol) servers should be disabled to avoid conflicts [8].

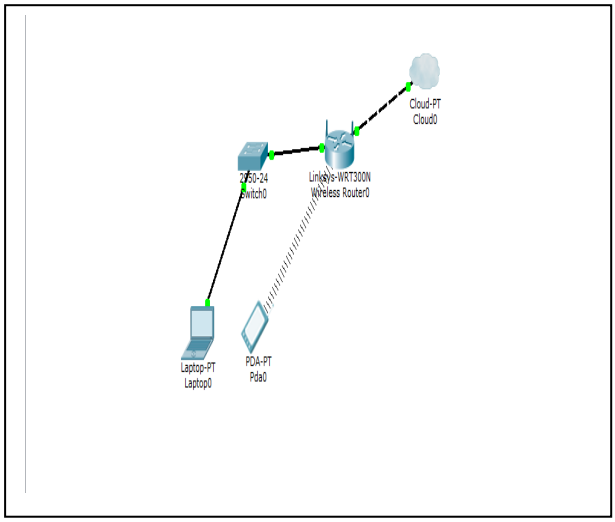


Figure 3. Firewall in the network

- Web Filter is a program that can screen an incoming web page to determine whether some or all of it should be displayed to the user. It allows enterprise to block out pages from the websites. This can be enabled in a school to prevent the end-users from having access to objectionable advertising, pornographic content, spyware and viruses.[6] Some web filters also provide reporting so that the installer can see what kind of traffic is being filtered and who has requested it. It is often installed as part of a proxy server and firewall.
- VTP (VLAN Trunking Protocol) reduces administration in a switched network, when you configure a new VLAN on one VTP server, the VLAN is distributed through all switches in the domain. This reduces the need to configure the same VLAN everywhere VTP is a cisco proprietary protocol that is available on most of the cisco catalyst series product.
- Port Security is used when MAC addresses are dynamic or static which reduces the MAC addresses that are allowed to send traffic into the port. Therefore, if the number of secure MAC address is one, and is assigned to the port, only the device attached to that port has full bandwidth of that port. This can reduce the risk of having unnecessary devices gaining access to the network. Having the number of devices that can access a port can be utilized by making a maximum of 2 mac-addresses. This is suitable for a classroom considering how

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

different students can have different lectures in the same room.

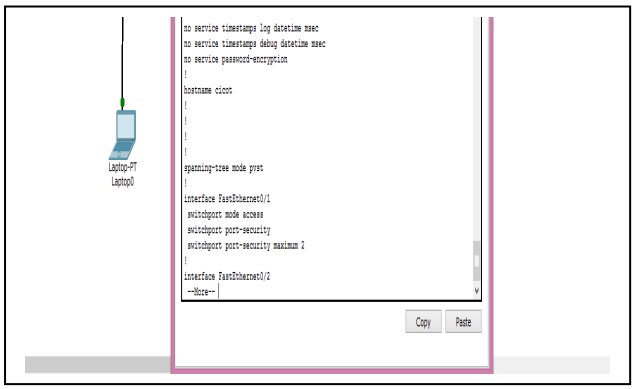


Figure 4. Configuring mac-address security

IV. RECOMMENDATION

To reduce threats on a network requires a lot of dedication and continuity. Increase In security of a network by using all the methods mentioned above can result in a standard and efficient production of services to end-users. There is no such thing as a "too secured" network, and so the environment should give room for all these measures. This is made easy by using NAC majorly because of its interoperability and flexibility in any enterprise of use. NAC is best used with cisco devices to achieve a very reliable system using ideal protocols and obtain the following results;

- Larger broadcast domains can lead to network failures. The use on VLAN helps reduces such challenges and gives way for multicast which is a foundation for improved broadcast and wastage for reliable LANs.
- The use of PVST on a network prevents the attack on one VLAN from affecting other VLANs.
- It increases network stability, dynamic learning which is unrealistic with the existing network.
- NAC with a cisco architecture allows students and staffs have their devices connect safely to the internet.
- Logging: enables administrators have a detail of the activities/attacks on the network.
- Policies that guide the use of devices and accessing of resources in a network can be configured during pre-admission or post-admission to curb security threats.
- Improved network performance as there is guaranteed network resources protection from activities on the internet.
- Improved data integrity of the institution's files.

CONCLUSION

The case study is an environment with a need for improved network access and also securing the network at large. Due to the region of the school, some modes of security employed in advanced settings cannot be put into practice considering low bandwidth signals from the ISP (Internet Service Provider), cost of maintenance and unequipped classrooms. Putting these protocols and services into action can kick-start a journey with solid foundation on a better network by using NAC with Cisco devices to enable continuity and a high level of maintenance. Thereof, as technology is evolving by the day, so is the need to get access to them for reasonable productivity.

REFERENCES

[1] D. Reed, Network model to information security, Nov 2003.

[2] M. Dark, R. Epstein, L. Morales, T. Countermine, Q. Yuan, Mohammed Ali, M. Rose & N. Harter, "A framework for information security ethics education", 10th Colloguium for information systems security education-University of Maryland.

[3] Jim Carr, Fast growing threats, SC Magazine US,IT security news, 2007

[4] John Adam, Data security, IEEE spectrum, v.29 n.8 p.19-20, August 1992.

[5] Network Access Control: Adapting security for a changing world, dimension data,2011

[6] Searchsecurity.techtarget.com/definition/web-filter

[7] www.cisco.com/c/en/us/support/docs/lan-switchong/spanning-tree-protocol/5234.5.html

[8] en.m.wikipedia.org/wiki/firewall_[computing]

AUTHORS PROFILE

Ayangbekun, Oluwafemi J received his Bachelor of Technology (BTech) in Computer Engineering from Ladoke Akintola University of Technology Ogbomoso, Nigeria in 2003. He also obtained his Masters of Science (MSc) in Computer Science from University of Ibadan, Nigeria in 2007. He is presently a PhD researcher in the Department of Information Systems, University of Capetown, South Africa

Audu, Zainab O received her Bachelor of Science degree in Computer Science with Statistics from Crescent University Abeokuta Ogun State Nigeria in 2015. Her area of interest includes Computer Network Security

A Survey: Multimodal Systems of Finger vein and Iris

Priyam Kaur Sandhu
CSE Department
PEC University of Technology,
Chandigarh, India

Manvjeet Kaur
CSE Department
PEC University of Technology,
Chandigarh, India

Abstract— **Biometric systems are the systems those enable automatic individual recognition which is based on behavioral or physical features belonging to a specific individual. All the biometric features have their limits to an extent and no biometric system is flawless. Therefore, the unimodal biometric systems have a lot of drawbacks. To solve the mentioned inconvenience and limitations and to enhance the level of security the multimodal biometric systems are employed. Personal identification process is a very important process that resides a large portion of daily usages. Human is a rich subject having many features that can be used for identification purpose such as finger vein, iris, and face etc. Finger vein recognition is an encouraging biometric recognition technique; the verification of individuals is done on the basis of the vein patterns present in the fingers. Iris recognition has moved under more focus due to its high reliability and efficiency in personal identification in past few years. This paper discusses the advantages of multimodal biometric system over unimodal biometric system along with the fusion of traits using different techniques. It also discusses the techniques employed on finger vein and iris. Finally, future scope is presented for the possible work that can be done in the field of finger vein and iris.**

Keywords- finger vein; iris; PCA; multimodal; unimodal

I. INTRODUCTION

A wide range of applications and systems need to confirm the identity of an individual before providing them services such as accessing a laptop, entering into a high alert area, ATMs, etc. Therefore it becomes really important to establish robust and efficient security systems to avoid breaches and impersonations. The customer and the industry fall into the trap of imposters due to the lack of effective customer verification. The traditional systems such as token based or knowledge based systems have limited security effects, up to a certain range these are secure, but beyond it these techniques are easy to hack or steal. The emergence of biometrics helped to address these issues. Biometrics refers to the process of authentication of an individual based on his/her behavioral or physiological traits. Physiological characteristics refer to the structure and shape of the body. Examples include finger vein, face, ear, DNA, palm print, iris, hand geometry, retina, etc. Behavioral characteristics are related to the manner in which a person behaves or shows his/her behavior, like voice, typing rhythm, gait, etc. They have an advantage over traditional security methods as they cannot be easily stolen or forged [1]. Initially a single trait was used for the authentication system called the unimodal biometric authentication system. As a single trait was used, so it possessed a number of downfalls like spoof attacks, non-universality, invariance, errors, etc. The limitations of unimodal biometric systems can be over powered by including numerous instances of identity of the same individual. The issue of non-universality is solved by using multiple traits that ensures coverage of population. Spoofing is taken care of as it would be difficult to forge the multiple biometric traits simultaneously. There are a lot of biometric traits which are used for authentication like iris, finger print, finger vein, face, gait, ear, DNA, hand geometry, voice, etc. Finger vein and iris are the two biometric traits which are the internal organs of the human body and are impossible to forge due to its unique characteristics.

II. MULTIMODAL

Unimodal biometric systems implement individual recognition on the basis of a single instance of biometric data. These systems have following limitations:

- It is easy to spoof, the data can be imitated or forged, e.g. rubber fingerprints can be used for spoofing
- Lack of permanence, with age the human traits change
- Cloning of a biometric characteristic
- Not being compatible with certain population
- Lack of universality
- Wear and tear of instruments causes error.

Multimodal biometric system refers to the system using a combination of two or more than two biometric modalities in the identification process. The most important reason to use the combination of different modalities is to enhance the recognition rate of the system. Therefore it is required that the feature sets of various biometric modalities are statistically independent. Anouar Ben Khalifa et al. proposed the fusion of two separate Unimodal biometric authentication systems. The fusion of the traits was made at different levels; for example feature level, sensor level, decision level and match score level. It includes the details about feature level fusion. In this case, fusion was carried out in parameters stage only when their feature sets were compatible. After the concatenation of the individual feature vectors, as a resultant, a single feature vector was obtained. In this case, the information was fused at the feature extraction level. The experimental results showed

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

that the feature level fusion performed better from rest [2]. Arun Ross and Anil K. Jain gave a comparative study between unimodal and multimodal biometric system. According to their research, unimodal biometric systems had plenty of issues such as intra class variations, the data had noise, degree of freedom was restricted, and spoof attacks were common, non-universality problem and unacceptable errors. Some of these limitations can be addressed by deploying multimodal biometric system that integrated the evidence presented by multiple source of information [3]. Arun Ross described the difference between biometrics and multibiometric with discussing the advantage of latter over former. It enumerated various sources of biometric information that can be consolidated to form a multimodal biometric system as well as the different levels of fusion in a biometric system. Along with the use of regular biometric traits used in designing a biometric system, this paper discusses the role of using ancillary information such as biometric data quality and soft biometric traits to enhance the performance of these systems [4]. Lindsay I Smith gave the detailed description of PCA algorithm along with its mathematical concepts and derivations. PCA is a useful statistical technique that has found wide range of applications in fields such image compression and face recognition. It was also used to find patterns in data of high dimension [5].

A lot of research has been done on the combinations of different traits in biometrics. A lot of multimodal systems are proposed in order to increase the efficiency of authentication and provide more security. Shruthi B.M et al. proposed a technique to combine Fingerprint and Finger Vein images using score level combination techniques i.e., holistic and nonlinear fusion. The results indicated that nonlinear approach performed better than average, product, weighted sum and likelihood ratio approaches [6]. Yongming Yang et al. presented the features of fingerprint and finger vein by minutiae which were compatible in nature. The quality of features at specific interested areas was evaluated for accurate matching. To weaken the influence of low quality images and false features, a dynamic weighting strategy was explored based on the results of feature evaluation. This scheme achieved 98.9% recognition rate. Dynamic weighting algorithm can achieve better performance even though poor quality fingerprint images were presented [7]. Kang et al. proposed a multimodal system integrating finger geometry and finger vein recognition at the score level fusion. The multimodal of both the traits was constructed in one device. The finger geometry recognition was based on Fourier descriptors; it was robust and managed the rotation of a finger. The equal error rate of the proposed method decreased by 1.089 [8]. Chen, Ching-Han, and Chia Te Chu combined face and iris features for developing a multimode biometric approach. The results showed that the multimodal biometric system was more reliable and precise as compared to the single biometric system [9]. Ko, Teddy presented various conditions those were possible in multimodal biometric systems using the traits: fingerprint, face and iris recognition. The various levels of fusion and the integration strategies that could be implemented to fuse information and enhance the overall system accuracy were also discussed [10]. Heng Fui Liau suggested a face and iris multimodal biometric system based on match score level fusion. In this process a support vector machine (SVM) was used. The performances of iris and face recognition systems were enhanced with the implementation of proposed feature selection method. SVM was used for computational speed-up. The results showed increased accuracy in terms of total error rate [11]. Besbes proposed a multimodal biometric authentication system based on fingerprint and iris biometric traits [12]. Byung Jun Kang proposed a new recognition method using finger, the technique was based on the score level fusion of the traits: finger veins, fingerprints and finger geometry features. The finger-vein and fingerprint recognition performances were enhanced by using local derivative pattern method. Fourier descriptor was combined with principal component analysis to increase the accuracy of the finger geometry recognition. All the three traits were combined with the help of three support vector machines and a weighted SUM rule. The outcomes showed that the equal error rate of the proposed method was accounted to be 0.254% in the experimental process [13]. Wenming Yang fused finger vein pattern with finger-dorsa texture. Initially, finger vein images and finger-dorsa images were captured simultaneously from the same finger. Then segmentation of the regions of interest of vein image and dorsal image were done. The features were extracted from finger-dorsa texture and finger vein pattern. Vein extraction method consisted of following activities: local thresholding was carried out, then modified line tracking was used, thorough probability map was created and directional neighbor analysis was done. To extract finger-dorsa texture, gray normalization was done on the finger-dorsa image. Then finally, the normalized dorsal texture and binarized vein patterns were fused together into one feature image [14].

III. BIOMETRIC TRAITS

A number of biometric traits are available for the authentication purpose. In this paper we will be focusing on the work done on finger vein and iris as these are the two traits those are highly secure being the internal organs of the body. Finger vein being the new trait in the market has not gathered much of spot light in application area and iris being one of the most trusted biometric traits is widely used.

A. Finger Vein

Finger vein based blood vessel patterns are unique for each individual. They have high level of security because the veins are located under the surface of the skin. Finger vein is emerging as a trusted biometric trait as the other traits like fingerprints can be cheated by dummy finger fitted with an imprinted fingerprint; on the other hand the finger vein based identification system is highly secure for authentication. Vascular scanners do not require the person to come in contact with the live scanners, so the person does not come in contact with any of the external devices. Since the information they read is present inside the body, the skin conditions do not affect the reading accuracy. The vascular scanners have enormous speed; they scan in less than a second. Its accuracy is high as it extracts the inner vein pattern from under the skin of human which is impossible to change or copy. The acquisition of image of finger vein with accuracy is one of the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

main concerns. Naoto Miura et al. suggested a method to extract the vein pattern in fingers at the time of fluctuation in LED intensity. The robustness and the tolerance of finger vein extraction to irregular luminance and noise in pattern extraction algorithm was tested. This method proved to be better than the existing conventional methods [15]. In another research the error rate was reduced to certain level. The image acquired from the infrared light contained the patterns of vein and irregular shading produced due to the thickness of muscles and bones. To overcome this drawback, line tracking technique was employed at various positions of the image. An equal error rate of 0.145% was attained in personal identification in this technique [16]. Jiang Hong and Cao Qubo presented a method of optical illumination and detection to be used in near infrared finger vein image acquisition system. Maximum curvature method was used for finger vein pattern extraction in the process. The results showed a FAR of 1.65% [17]. Darun Tang et al. worked on the stability of the finger vein pattern. The difference in the finger vein template of the same person was more of a problem as compared to the similarity of templates of different persons. To tackle with this problem, Occurrence Probability Matrix (OPM) was introduced. The higher stability areas contributed to the results more as compared to the low stability areas. In this way, the similarity of the template of the same person was evaluated and the influence of the unsteady areas were reduced. The results showed a drop of the EER of the system from 9.8% to 7.6% [18]. In another research Darun Tang et al. designed a template evolution method to reduce the effect of changing template when matching. On a long run, the captured templates could be different from the ones those were registered due to the user changing habits or device aging. Experimental results showed that by using this technique the equal error rate of the system got reduced from 6.3% to 3.5% [19]. Zhongbo Zhang et al. proposed a multiscale feature extraction method for finger-vein trait. This method was based on local interconnection structure neural networks and curvelets. The features of the finger-vein were extracted using neural network with local interconnection structure. The curvelets were used to execute the multiscale self-adaptive enhancement transform on the finger-vein images. The proposed method was superior to other conventional methods in finger-vein feature extraction. It extracted the features from obscure images in an efficient manner. The EER of the proposed method was experimentally evaluated to 0.128% [20]. Joon Hwan Choi et al. proposed a finger vein extraction method which included gradient normalization, binarization and principal curvature calculation. This method extracted the patterns of the finger vein regardless of the vein thickness or brightness [21]. MEI Cong-li et al. proposed a technique based on morphology. It improved the stability and robustness of the extracted finger vein patterns. The image boundaries were scanned and the valley detection was carried out from four directions. The feature extraction method improved the conventional point-wise comparison procedure. The average error rate was decreased with this technique [22]. WANG Ke-Jun YUAN Zhi presented the algorithm based on wavelet combined with Principal Component Analysis (PCA) transformation and LDA transformation. It eliminated the drawbacks of the single feature recognition and enhanced the speed of template matching. This technique provided fast and accurate identification [23]. Beining Huang et al. introduced a wide line detector for feature extraction of finger vein. The technique obtained information about the width of the vein and enhanced the information acquired from the extracted features from the low quality images. A pattern normalization model was also used based on elliptical cross sectional area of finger and closeness of the vein to the surface. Experimental results showed that the irregular distortions due to variance of finger pose got reduced [24]. Gongping Yang et al. extracted the features of finger vein using the technique $2D^2PCA$. A binary classifier was trained for each person based on metric learning. The KNN classifier was used for each individual. This technique was different as compared to the traditional methods in which a fixed threshold was applied for all individuals. The SMOTE technology was employed in order to solve the problem of class-imbalance. The proposed method attained a recognition rate of 99.17% [25].

B. Iris

The iris development takes place in the third month of life and unique patterns are formed during the first year of life. The patterns those are present in the iris of eye are unique to each individual [26]. Since the iris is an internal part of the body, so it is very secure and difficult to forge. Iris recognition systems are non-invasive in nature [27] [32], for practical applications it is a very important aspect. One of the most common and powerful tool used for such analysis is image processing techniques. Image processing can be used for formulation of an iris pattern to unique code which can be stored in a database and used for comparison purposes between templates and queries. In turn, the iris systems have a very low False Accept Rate (FAR) compared to other biometric traits that can be rather high. The feature extraction of iris is one of the main concerns in order to get secure authentication systems. Jong-Gook Ko et al. proposed a method for iris recognition. This method was based on iris feature extraction using a cumulative sum-based change analysis. The normalized iris image was divided into cells. For each cell the iris code was generated using proposed algorithm that used the cumulative sums of each cell. This method was efficient compared to existing methods. The proposed approach showed a good recognition performance and speed [28]. Kwanghyuk Bae et al. introduced a new algorithm for feature extraction of iris based on Independent Component Analysis. ICA was applied to generate the base vectors to extract efficient iris features. These vectors were localized in both frequency and space. The ICA expansion coefficients were used as feature vectors. Each of these vectors were used to generate iris code. Experimental results showed that the size of iris code and the time to process got significantly reduced [29]. Li Ma et al. proposed an algorithm in which the gabour filters were used to capture the global and local iris characteristics in order to form fixed length feature vector. Weighted Euclidean distance was evaluated between the two iris vectors for matching. This method increased the speed of matching process, extracted the features efficiently and was insensitive to illumination and noise [30]. Sunita V. Dhavale introduced a new technique for iris feature extraction based on Discrete Cosine Transform (DCT) domain. To detect the iris

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

boundaries in digital image, Canny Edge Detection and Hough Transform methods were used. After the preprocessing, two level Discrete Wavelet Transformation (DWT) was applied on the segmented images. To encode the iris features, both horizontal and vertical sub-bands were used. Each sub band was divided into blocks and DWT was applied on each of the blocks. The proposed technique proved to be computationally effective [31]. Wildes proposed an algorithm in which the image was converted into a binary edge map and Hough transform was used for circle detection. To extract features, Laplacian filter was used. Finally, normalized correlation was used for the matching between two iris images [32]. Mohammed Abdullah proposed an algorithm based on wavelet transform for iris recognition. The feature vector was stored in the form of binary code in this process [33]. Jing Huang et al. proposed an iris recognition system based on non separable wavelet. Initially, the iris image was decomposed into wavelet sub band coefficients with the help of sixteen non-separable wavelet filters. Then, Generalized Gaussian Density (GGD) modeling was used for feature extraction of each orthogonal wavelet coefficient. For the purpose of matching, Kullback Leiblar distance was computed between GGDs [34]. Liu Yang et al. used multi-scale 2D Gabor filter to attain the code of iris. In order to protect the code, one-way coupled map lattice (OCML) chaos system was applied that generated pseudo-random number key stream. Cipher text feedback was employed to encrypt and decrypt the iris code. Finally, Hamming distance was used for iris classification. The proposed system provided a high recognition rate and high encryption speed [35]. Zhonghua Lin and Bibo Lu used the imaginary coefficients of Morlet Wavelet Transform to generate the binary code of the iris images at different scales [36]. Xu Xiuli et.al gave an approach to obtain the effective iris feature matrices with lower dimension. The feature extraction method used was Complete Two-Dimension Principal Component Analysis. C-2DPCA performed better than both 2DLDA and 2DPCA with a lower Equal Error Rate (EER) and an average computation time [37]. Attarchi et al. used a complex mapping procedure and best-fitting line for the iris segmentation and ID Gabor filter with 2DPCA for the recognition approach. An intensity threshold method with canny edge detector was used to extract the rough region of the pupil. To localize the rough region of the outer boundary, the method of median filter having prewitt compass edge detector was employed. The bottom point of the pupil was selected as a reference point. The two sets of intersecting points between the horizontal lines and pupil's inner and outer boundaries were created with respect to the reference. Each point set was then mapped into a new complex domain using the complex inversion map function and the best-fitting line was found on the range. By remapping the best-fitting lines to the original domain, the exact inner and outer boundaries of the iris were found. In order to reduce the dimensionality of the extracted features, 2DPCA method was used [38].

IV. Future Work

Finger vein and iris biometric systems are very promising authentication systems. As both are internal parts of the body, so it is impossible to forge the templates. Finger vein and Iris individually have given great results in authentication of individuals. A lot of work has been done on the fusion of finger vein and iris with other biometric traits using sensor level, match score and decision level fusion. The combination of finger vein and iris is still not common. The fusion of finger vein and iris using feature level fusion can be done based on Principle Component Analysis technique. The combination is unique as very little work has been done on it. These traits have formed a perfect multimodal system and achieved high accuracy independently with different techniques with other biometric modalities. Therefore the implementation of two dimensional square Principal Component Analysis ($2D^2PCA$) on combination of finger vein and iris using feature level fusion could enhance the efficiency of the system.

References

[1] Ross, Arun, and Anil Jain. "Information fusion in biometrics." *Pattern recognition letters* 24.13 (2003): 2115-2125.

[2] Anouar Ben Khalifa, Najoua Essoukri Ben Amara, "Bimodal Biometric verification with different fusion levels", 6th International Multi-Conference on Systems, Signals and Devices, pp. 1-6, March 2009.

[3] Arun Ross and Anil K. Jain, "Multimodal Biometrics: An Overview", Appeared in Proc. of 12th European Signal Processing Conference (EUSIPCO), (Vienna, Austria), pp. 1221-1224, September 2004.

[4] Arun Ross, "An introduction to Multibiometric", European Signal Processing Conference (EUSIPCO), Poland, September 2007.

[5] Lindsay I Smith, "A tutorial on principal Component Analysis", Feb 2006.

[6] Shruthi B.M et al., "Multimodal Biometric Authentication Combining Finger Vein and Finger Print", International Journal of Engineering Research and Development, Volume 7, Issue 10 (July 2013), PP. 43-54

[7] Yongming Yang, et al., "Dynamic Weighting for Effective Fusion of Fingerprint and Finger Vein", Progress in Intelligent Computing and Applications, Volume1, Number1, October 2012

[8] Kang, Beyun J., and Kang R. Park. "Multimodal biometric method based on vein and geometry of a single finger." *Computer Vision, IET* 4.3 (2010): 209-217.

[9] Chen, Ching-Han, and Chia Te Chu. "Fusion of face and iris features for multimodal biometrics." *Advances in Biometrics*. Springer Berlin Heidelberg, 2005. 571-580.

[10] Ko, Teddy. "Multimodal biometric identification for large user population using fingerprint, face and iris recognition." *Applied Imagery and Pattern Recognition Workshop, 2005. Proceedings. 34th*. IEEE, 2005.

[11] Liau, Heng Fui, and Dino Isa. "Feature selection for support vector machine-based face-iris multimodal biometric system." *Expert Systems with Applications* 38.9 (2011): 11105-11111.

[12] Besbes, Feten, H. Trichili, and Basel Solaiman. "Multimodal biometric system based on fingerprint identification and iris recognition." *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*. IEEE, 2008.

[13] Kang, Byung Jun, et al. "Multimodal biometric method that combines veins, prints, and shape of a finger." *Optical Engineering* 50.1 (2011): 017201-017201.

[14] Yang, Wenming, Xiang Yu, and Qingmin Liao. "Personal authentication using finger vein pattern and finger-dorsa texture fusion." *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009.

[15] Miura, Naoto, Akio Nagasaka, and Takafumi Miyatake. "Automatic Feature Extraction from non-uniform Finger Vein Image and its Application to Personal Identification." *MVA*. 2002.

[16] Miura, Naoto, Akio Nagasaka, and Takafumi Miyatake. "Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification." *Machine Vision and Applications* 15.4 (2004): 194-203.

[17] Hong, Jiang, and Cao Qubo. "The finger vein image acquisition method and vein pattern extraction study based on near infrared." *World Automation Congress (WAC), 2012*. IEEE, 2012.

[18] Tang, Darun, et al. "Finger vein verification using Occurrence Probability Matrix (OPM)." *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012.

[19] Tang, Darun, et al. "A Method of Evolving Finger Vein Template." *Biometrics and Security Technologies (ISBAST), 2012 International Symposium on*. IEEE, 2012.

[20] Zhang, Zhongbo, Siliang Ma, and Xiao Han. "Multiscale feature extraction of finger-vein patterns based on curvelets and local interconnection structure neural network." *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 4. IEEE, 2006.

[21] Choi, Joon Hwan, et al. "Finger vein extraction using gradient normalization and principal curvature." *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009.

[22] Cong-li, Mei, et al. "Feature extraction of finger-vein image based on morphologic algorithm." *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*. Vol. 3. IEEE, 2009.

[23] Zhi, WANG Ke-Jun YUAN. "Finger Vein Recognition Based on Wavelet Moment Fused with PCA Transform [J]." *Pattern Recognition and Artificial Intelligence* 5 (2007).

[24] Huang, Beining, et al. "Finger-vein authentication based on wide line detector and pattern normalization." *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010.

[25] Gongping Yang, et al., "Finger Vein Recognition Based on $2D^2PCA$ and Metric Learning", Hindawi Publishing Corporation Journal of Biomedicine and Biotechnology, 2012

[26] H. Davision, The Eye, Academic, London, 1962.

[27] J.G. Daugman, "High Confidence Visual Recognition of Persons by a Test of Statistical Independence", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.15, no.11, pp.1148-1161, Nov. 1993.

[28] Ko, Jong-Gook, et al. "A novel and efficient feature extraction method for iris recognition." *ETRI journal* 29.3 (2007): 399-401.

[29] Bae, Kwanghyuk, Seungin Noh, and Jaihie Kim. "Iris feature extraction using independent component analysis." *Audio-and Video-Based Biometric Person Authentication*. Springer Berlin Heidelberg, 2003.

[30] Ma, Lia, Yunhong Wang, and Tieniu Tan. "Iris recognition based on multichannel Gabor filtering." *Proc. Fifth Asian Conf. Computer Vision*. Vol. 1. 2002.

[31] Dhavale, Sunita V. "DWT and DCT based robust iris feature extraction and recognition algorithm for biometric personal identification." *International journal of computer applications* 40.7 (2012): 33-37.

[32] Wildes, Richard P. "Iris recognition: an emerging biometric technology."*Proceedings of the IEEE* 85.9 (1997): 1348-1363.

[33] Abdullah, Mohammed AM, et al. "Smart card with iris recognition for high security access environment." *Biomedical Engineering (MECBME), 2011 1st Middle East Conference on*. IEEE, 2011.

[34] Huang, Jing, Xinge You, and Yuan Yan Tang. "Iris recognition based on non-separable wavelet." *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*. IEEE, 2008.

[35] Yang, Liu, et al. "Iris recognition system based on chaos encryption."*Computer Design and Applications (ICCDA), 2010 International Conference on*. Vol. 1. IEEE, 2010.

[36] Lin, Zhonghua, and Bibo Lu. "Iris recognition method based on the imaginary coefficients of Morlet wavelet transform." *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*. Vol. 2. IEEE, 2010.

[37] Xu, Xiuli and Ping Guo, "Iris feature extraction based on the complete 2DPCA", Advances in Neural Networks–ISNN 2009, Springer Berlin Heidelberg, 2009, 950-958.

[38] Attarchi, Sepehr, Karim Faez and Amin Asghari, "A fast and accurate iris recognition method using the complex inversion map and 2DPCA", Computer and Information Science, 2008, ICIS 08, Seventh IEEE/ACIS International Conference on, IEEE, 2008.

Embedded Mobile Agent (EMA) for Distributed Information Retrieval

Oguntunde, B.O
Department of Computer Science
Redeemer's University,
Ede, Osun State, Nigeria

Osofisan A.O
Department of Computer Science
University of Ibadan, Ibadan,
Oyo State, Nigeria.

Aderounmu, G.A
Department of Computer Science and Engineering
Obafemi Awolowo university,
Ile-ife, Osun State, Nigeria

***Abstract*— Mobile agent paradigm has been recognised as a viable approach for building distributed applications. Mobile agents migrate through the network, execute asynchronously and autonomously, conserve bandwidth, achieve better load balancing, adapt dynamically to changes in their environment, are robust and fault tolerant. Existing agents run and execute on agent platforms also called, the Mobile Agent System (MAS), which provides run-time execution and support facilitites for mobile agent to accomplish it tasks. These MASs from different vendors are different in laguage, design,and implementation and are not interoperateable, this impedes the achievement of the full potentials of mobile agent paradigm. This work is aimed at providing a robust structure for deploying mobile agents so they can execute independent of the MAS. We propose a lightweight agent to run in the kernel mode of the operating system as an operating system service, giving an impression of the agent directly communicating with the operating systems.**

.Keywords- embedded mobile agent, operating system service, lightweight agent, agent platform.

I. INTRODUCTION *(HEADING 1)*

Mobile agent paradigm has been recognized as a viable tool and a promising approach for building distributed applications [1, 2, 3] and a lot of research has been done, nevertheless, it is still a promising area of research, because, a lot of its many potentials are yet to be exploited. Agents solve complex software problems in distributed environments where protocols, operating systems, hardware and runtime environments are heterogeneous. Mobile agent has been defined as a computer entity capable of reasoning, use the network infrastructure to run in another remote site, search and gather the results, cooperate with other sites and return to its home site after completing the assigned tasks [4]. Mobile agents system provides infrastructure for executing automous agents and also migrate them between computers connected by a network.

Mobile agent paradigm was proposed as an alternative to client server [1], it offers flexibility on the reliance on network connection [2]. Once launched, can be disconnected, it keeps performing its tasks and can be reconnected to receive the result at a later time [5]. Mobile agents have the potentials to improve the speed and efficiency of computation by moving computation to data [6], thus eliminating unnecessary and massive data transfer over the network. According to [7], they are viable tools when information needed is vast and widely distributed, and in application or service that needs to learn and improve over time.

Mobile agent technology consists of mobile agents and mobile agent middleware also called platform (MAS). Mobile agent platform is a distributed execution environment for mobile agents [8], MAS provides services and primitives that help in the use, implementation and execution of system development using mobile agent paradigm [9]. Mobile agents run and execute on mobile agent platforms that provide run-time execution for mobile agents. The platforms are installed on the computers in the systems on which the agents are expected to run which consumes memory, increases access time and prevents other tasks from being run on the computer. There are many different agent platforms developed to support agent applications [10, 11, 12] and these platforms are not interoperateable, i.e an agent built on one platform cannot run on another platform. Mobile agents are naturally heterogeneous and should not be limited by agent platforms. This is one of the issues impeding the global acceptance of mobile agent paradigm as the absolute solution for the next generation distributed systems. There is therefore, a need for a unified system to run and execute mobile agents regardless of the design, platforms and vendor. Efforts made by the Foundation for Intelligent and Physical Agent (FIPA) to achieve interoperability among agents and Mobile Agent System Interoperability Framework (MASIF) to achieve interoperability among MASs [13] are yet to be fully achieved [14]. Furthermore, mobile agents have enormous potentials that are yet to be discovered, research in the past has focused mainly on the application of this technology to solve complex problems or improve certain solutions to sophisticated software problems, but little is being done on the improvement of the technology itself. Improving the mobile agent technology is the major focus of this work.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

In retrospect, mobile agent has been applied in many areas of research such as, information retrieval and management [15, 16], electronic commerce [17, 18], grid job scheduling [19], expert finding [20], network management [1, 21], traffic detection and management [22], examination system [23], supply chain management [24] and many more. A lot of issues arose with the use of mobile agents among which are security, complexity and lack of standard [25, 26]. The complexity and sophistication naturally led to many attempts to simplify and extend agent functionality, thus attention shifted to providing necessary security for mobile agents, agent platforms and hosts on which they execute [27]. The versatility of mobile agent paradigm also increased research interest in enhancing mobile agents in the area of agent communication [28] and agent structure [12], in order to extend their functionalities. It is on this note that we attempt to enhance the mode of deployment of mobile agents in order to make them execute without going through the agent platform. This work, attempts to eliminate agent platforms and make agent run as part of the hosts' operating system.

This work presents an embedded mobile agent that offers the possibility of mobile agents interacting directly with the operating systems on the host computers. The existing mobile agents require agent platforms to be previously installed on the computers on which they are to run; this platform needs to be explicitly initiated before receiving and providing runtime execution for incoming mobile agents. This work takes advantage of the fact that all computers run an operating system and attempts to make agents part of the operating system. The kernel mode of the operating system (specifically Windows Operating System) is extended with a static agent as a service in the kernel mode. The focus of this research work is to eliminate the agent platforms and make mobile agents run as part of Windows O/S in form of operating system service.

II. WINDOWS OPERATING SYSTEM SERVICE

Windows Operating system provides a way of making certain programs available to run as part of the operating system in the form of operating system service. Operating system Service is a long running executable program that runs in its own windows session, without user's intervention, this is similar to daemon in Unix operating systems [29].

The existing mobile agent for information retrieval model consists of mobile agents that execute on agent platforms previously installed on the computer machine. The platform is installed in memory on top of the operating system running on the host, this obviously consumes memory, and increases access time. The framework provided in this work includes a light weight agent embedded into the kernel mode of the operating systems to free memory and reduce access time. Figure 1 presents the proposed embedded agent as Windows XP operating system service. A light weight static agent that receives and executes mobile agents was embedded into the kernel mode of the Windows XP operating system as part of the executive services.

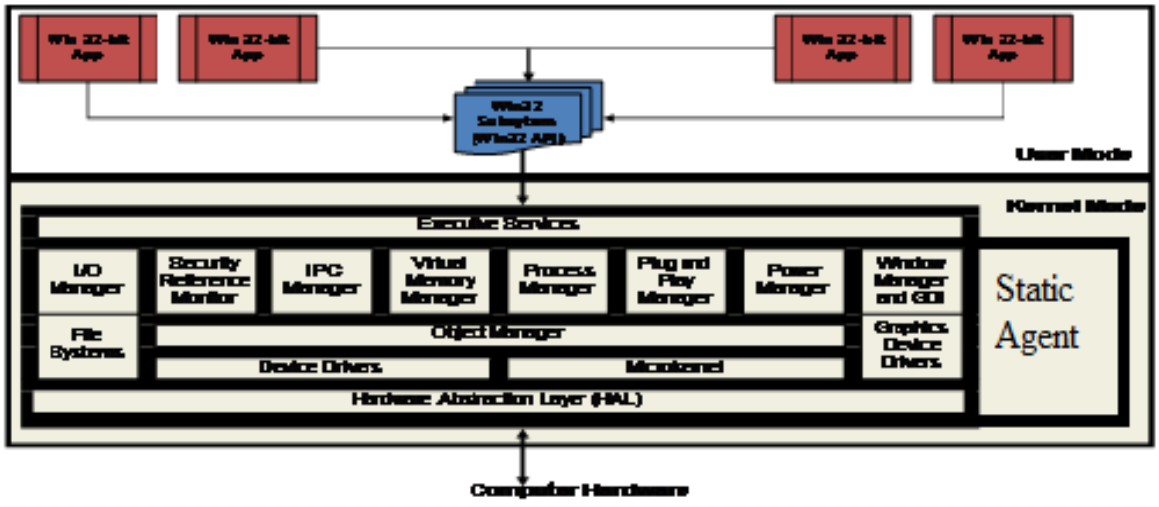


Figure 1: embedded agent as Windows XP operating system service (Adapted from 30)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

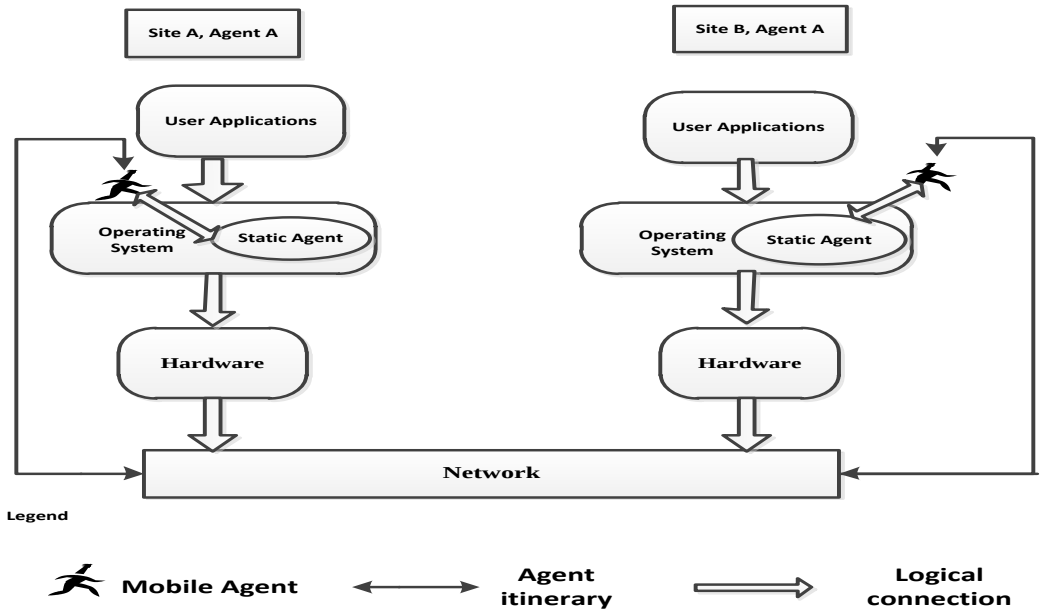


Figure 3: the conceptual architecture of the proposed system.

The embedded agent was designed using the layered architecture, such as the three layer architecture designed by [24]. Each layer represents a particular function. A mobile agent contains code, state information and attributes. The attributes of mobile agent include its name which is unique for identification, the authority or the owner of the agent and the agent system type. Code contains the logic of the agent, i.e. code defines the behavior or the required tasks of the agent, same type of agents use same code. Code in an object oriented context means the class code necessary for an agent execution [31]. Data corresponds to the value of the agent's instance variables and include information about the mobile agent such as its launcher, movement history, resource requirements and authentication keys for use by the infrastructure; these are referred to as the initial data. The data also include results of the mobile agent's tasks on different nodes visited, referred to as the generated or received data.

The enhancement is a static agent embedded and made to run in the kernel mode of the operating system as an Operating System Service. The static agent receives and provides execution environment for visiting mobile agents as depicted in figure 2, this gives the impression of mobile agents communicating directly with the operating system. The target operating system is the Windows OS (Windows XP, Windows vista and Windows 7). The mobile agent class has attributes execute state, communicate, record and the list of tasks to perform. Figure 2 shows the conceptual architecture of the proposed system.

III. IMPLEMENTATION

To create a customised service involves setting up the inheritance and other infrastructure elements. The static agent class inherits from the ServiceBase class and a main method defines the service to run, the following implements the static agent

```
public class StaticAgent implements Runnable{
    private Socket insoc;
    private ObjectInputStream ois;
    public static int AGENTS_PORT = 4999;
    public StaticAgent(Socket insoc) throws IOException
    {
        this.insoc = insoc;
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
        // ……..;
    }
    public void run() {
        try{
                MobileAgent magent =
(MobileAgent) ois.readObject();
//                ……..
        System.out.println(ex);
             }finally{
                try{
            }catch(Exception x){}
         }
    }.
```

We evaluate the proposed Embedded Mobile Agent (EMA) architecture in distributed information retrieval environment, Let's consider the situation at the meteorological agency with headquarters in Abuja, Nigeria and three other branches each representing one geopolitical zone, say, Lagos for southwest, Port-Harcourt for Eastern Zone and Kano for the northern Zone out of the six geopolitical zones in Nigeria. Each branch agency runs an embedded agent as part of their operating system, provides execution environment with heterogeneous hardware configurations and versions of Windows operating system (Windows XP, Windows Vista and Windows 7).

using the case of retrieving weather information. The information was stored in databases that are distributed geographically and connected by a network. Four hypothetical locations were chosen and connected together by a network as depicted by figure 3, the Mobile Agent was sent from one location given the itinerary and it visited the other nodes collected the required information and returned to the origin. On each host, the static agent was installed and ran continuously, without users intervention, it is lunched automatically once the computer boots. The static agent listens to the port for incoming mobile agents, negotiates passage to the destination host, validates and authenticates the incoming agent, launches the received mobile agent and provides runtime execution for the agent. The mobile agent migrates through the network, negotiates access with the static agent on remote host, downloads the required information and adds it to its bag. It then determines the next node to visit, initiates a move to the next node, returns to the origin with the result of the search and disposes itself.

The director at the headquarters in Abuja (origin) requires the weather conditions (temperature and atmospheric conditions) from the different stations in each zone.
The EMA is dispatched from the origin (Abuja) given the itinerary (names or IP addresses) of the nodes to visits. The EMA migrates to first node in its itinerary, retrieves the required information and migrates to the next location in its itinerary, repeats the same process and to the next performs its function and later returns to the origin with the result of the search. To retrieve information on each host, the EMA performs SQL queries on relational databases which contain records in a table with structure described by figure 4.

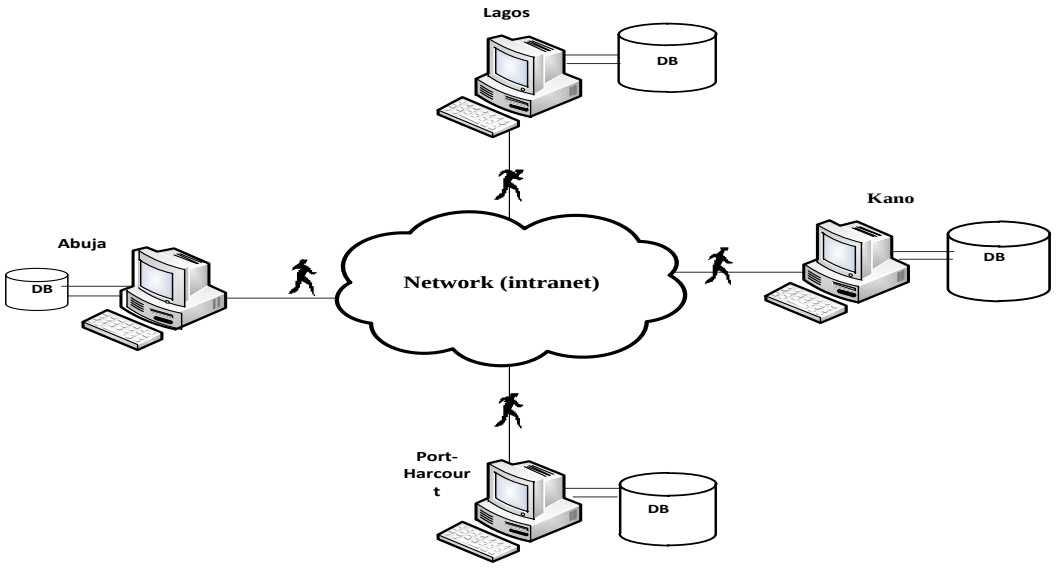


Figure 3: Overall architecture of the proposed system

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

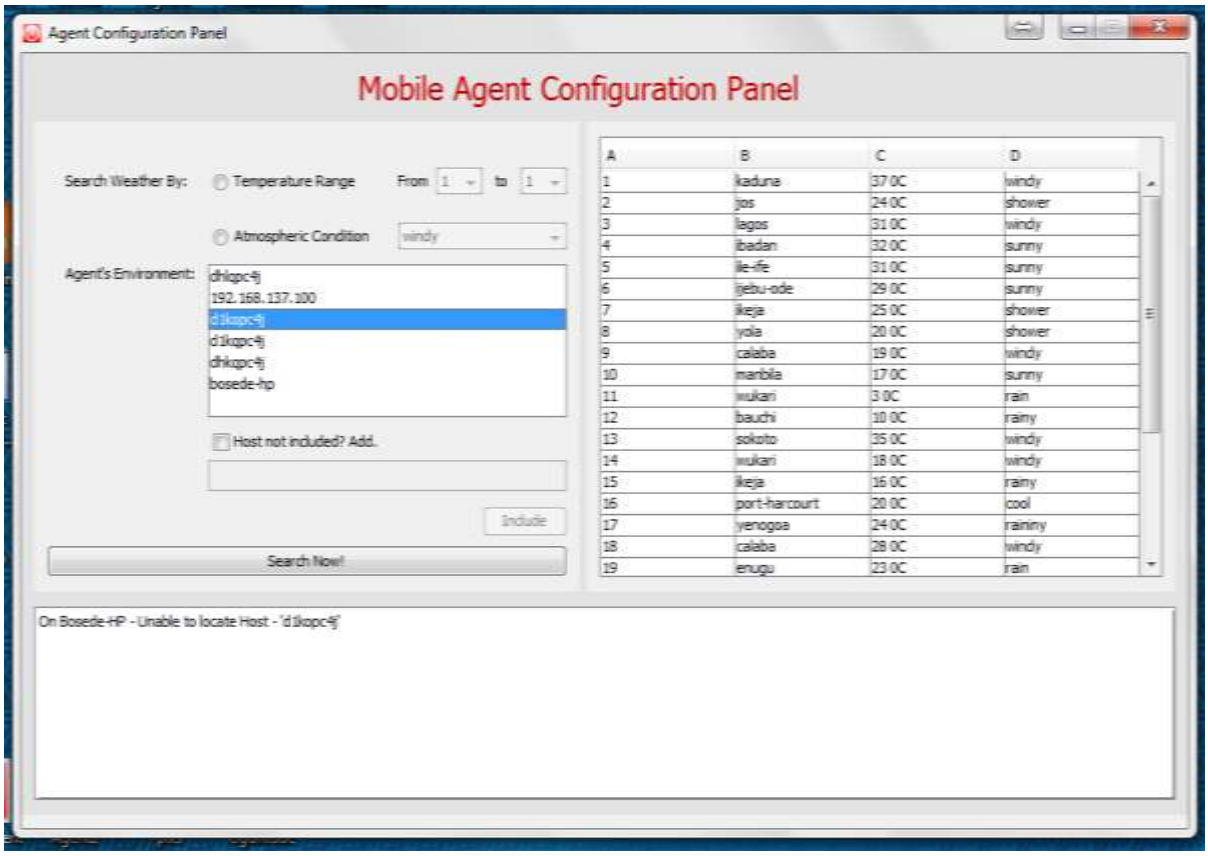


Figure 4: Search sample with the EMA

Experimental results

EMA was deployed into an existing local area network and its performance was compared with that of existing Java Agent Development Framework (JADE). The EMA proves to be a more efficient and autonomous scheme with a high level of flexibility compared to JADE. It reduces memory consumption, reduces access time, is robust and fault tolerant.

Conclusion

In this work, we present Embedded Mobile Agent (EMA) architecture, based on the common mobile agent structure with an additional feature added. The system provided by [3] also focussed on enhancing the structure of mobile agent to improve future performance while executing on agent platform. Whereas, the system presented in this work was designed to be embedded in the Windows Operating System in the form of Windows service to enable mobile agents directly interacts with the Operating System. The implementation of the Embedded Mobile Agent was achieved and deployed into an existing Local Area Network to retrieve information from remote distributed databases. The potentials of mobile agents can be extended by eliminating the barrier placed by agent platforms.

In the area of future research, implementation with other operating systems such as Unix or any of its flavours will be investigated and the EMA will be applied to more sophisticated distributed operations. In addition, we intend to provide adequate security for the Embedded Mobile Agent.

References

[1] G.A. Aderounmu, "Development of an intelligent mobile agent for computer network performance management," Ph.D dissertation, Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria, 2001

[2] G.A. Aderounmu, B.O. Oyatokun and M.O. Adigun. Remote Method Invocation and Mobile Agent: a Comparative Analysis. *Issues in Informing Science and Information Technology,* vol. 3, 2006, Available: http://informingscience.org/proceedings.INSITE2006/IISTAder188.pdf

[3] S. Gabriel and I. P. Claudiu, "A Proposal for an Enhanced Mobile Agent Architecture (EMA)". *Annals of the University of Craiova, Mathematics and Computer Science Series. 37(1): 71 – 79, 2009*

[4] O. Abdelkader, "Mobile Agent-Based Applications: A Survey," *International Journal of Computer Science and Network Security, 331 – 339, 2009.*

[5] J. Dale and D.C. DeRoure, "A Mobile Agent Architecture for Distributed Information Management", *in proc. 1997 International workshop on the virtual Multicomputer, 1997.*

[6] G.P. Picco,. "Understanding code mobility". Technical report, Dipartimento di Elettronica Informazione, Politecnico di Milano, 2005. [date of last access: 18 June, 2013]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] I. Sridhar, and J. Vikram, "Designing Distributed Applications using Mobile Agent," in pro*c. 2001 International Conference on High Performance Computing, 2001. Hyderabad, India.*

[8] S. Dilyana and G. Petya, "Building Distributed Applications with Java Mobile Agent", presented at the International workshop NGNT, 2002.

[9] S. S. F. Roberto, "The Mobile Agents Paradigm," research paper, department of Information and Computer Science, University of Califonia, Irvine, 2001.

[10] A. Syed, D. John and Y. Pavana, "A survey of mobile agent systems," Student Report, Department of Computer Science and Engineering, University of California San Diego, 2000. (Date of last access: 03 July, 2013 available cseweb.ucsd.edu/classes/sp00/cse221/reports/dat-yal-and.pdf).

[11] G. J. van 't Noordende, B. J. Overeinder, R. J Timmer, F. M. T. Brazier and A. S. Tanenbaum, 2009. "*Constructing secure mobile agent systems using the agent operating system"*, International Journal of Intelligent Information and Database Systems, Vol. 3, No. 4,

[12] Bellifemine, F.L, Greenwood D and Caire G. 2007. *Developing Multi-agent systems with JADE.* John Wiley & Sons Ltd, England.

[13] Roth V., Pinsdof U. and Binder W. Mobile agent interoperability revisited. Available: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.1814&rep=rep1&type=pdf

[14] Clark K. L and Lazarou V. S. 1997. A Multi-Agent System for Distributed Information retrieval on the World Wide Web.

[15] Htoon H. and Mie M. T. T. 2008. Mobile Agent for Distributed Information Retrieval System. *Proc. 2008 ECTI-CON.*

[16] Brian B., Robert G., Katsuhiro M., David k., George C. and Daneila R. 1999. Mobile Agent in Distributed Information Retrieval. Technical report submitted to Thayer School of Engineering, department of Computer Science Dartmouth College Hanover, new Hampshire. Firstname.lastname@dartmouth.edu.

[17] Christoph B., Volker R. and Ralph M. 2002. Perspectives on Electronic Commerce with Mobile Agents.

[18] Enock M. 2005. A Development of Resource/Commander Agents Used in AgentTeamwork Grid Computing Middleware.

[19] Iyilade, J.O, Aderounmu G.A and Adigun, M.O. 2005. An Agent-based Approach for finding a Supervisor in an Academic Environment. *Proceedings of the 3rd International Conference on Education and Information Systems: Technologies and Application. Orlando, Florida, USA, 363–381.*

[20] Bohoris C. (2003). Network Performance Management Using Mobile Software Agents. Ph. D thesis, University of Surrey, UK.

[21] Bo Chen, Henry H. C. and Joe P. 2009. Integrating mobile agent technology with multi-agent systems for distributed traffic detection and management systems. Transportation Research Part C, pp 1-10.

[22] Gawali, R.D and Meshram, B.B. 2009. Agent-Based Autonomous Examination Systems. *Proceedings of the Intelligent Agent and Multi-agent systems, 2009, (IAMA 2009) International Conference.*

[23] Wenjuan W., Tong Li, Weidong Z. and Weihui D. 2009. Mobile agent system for supply chain management. *Proceedings of the second symposium of International Computer Science and Computational technology (ISCSCT'09), Huangshan, P.R China, 525-528.*

[24] Fortino G. and Russo W. 2003. High-level interoperability between java-based mobile agent systems. A report of the project ' Giovane Ricercatore 2003', University of Calabria.

[25] Tudor M., Bogdan D., Mihaela D. and Ioan S.. A Framework of Reusable Structures for Mobile agent Development. *Proceedings of the 8th IEEE international Conference on Intelligent Engineering Systems (INES '04)*, Cluj-Napoca, 2004.

[26] Biermann E. 2004. A Framework for the Protection of Mobile Agents Against Malicious Hosts. A Ph.D thesis, University of South Africa, South Africa.

[27] Priya, B.G; Suba, S.; Bensal, T.; and Boominathan, P. 2009. Enhanced Communication Scheme for Mobile Agent, *Proceedings of the Intelligent Agent and Multi-agent systems, 2009, (IAMA 2009) International Conference.*

[28] Wikipedia 2013. Windows Service. Available: www.en.wikipedia.org/wiki/Windows_service. [date of las access: 24 July, 2013]

[29] WIN133. (2009). Introduction to Windows XP architecture. Available: www.warrenworks.com/csc_110/SupportFiles/w2karchitecture.ppt. [Retrieved on 20 *September*, 2012.

[30] Lange D. B. And Oshima M. 1999. Seven Good Reasons for Mobile Agents. *Communication of the ACM, .42(3).*

AUTHORS PROFILE

Oguntunde, B.O holds a B.Tech (Hons) degree in Computer Engineering from LAUTECH, Ogbomosho, Nigeria, in 2000. She obtained her M.Sc and PhD in computer Science from the University of Ibadan. She is currently teaches at the Redeemer's University, Nigeria. Her research interests are in the areas of computer communication and deployment of mobile agent software management of heterogeneous computer network.

Osofisan, A.O. obtained a B.Sc (Hons) degree in Computer Science from the Obafemi Awolowo University, Ile Ife, M.Sc from Georgia Tech. and PhD from Obafemi Awolowo University. She is currently a Professor and the director of Business School, The University of Ibadan where she also lectures in the department of Computer science. Her areas of specializations are data communications, data warehousing and data mining. She has many articles in these areas at both local and international level to her credit.

Aderounmu G.A. obtained a obtained a B.Sc. (Hons) degree in Computer Engineering from the Obafemi Awolowo University, Ile-Ife, Nigeria, in 1991. He obtained his M.Sc and Ph.D in Computer Science from the same University in 1997 and 2001 respectively. He lectures in the department of Computer Science and Engineering; he is presently a professor and the director of the Information Technology and Communication Unit (INTECU) of the same University. His areas of specializations are design, analysis, and simulation of ATM networks with respect to switching, protocol, and buffer management and mobile agent software development. He has many articles at both local and international level to his credit.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

FRAUDULENT ELECTRONIC TRANSACTION DETECTION USING DYNAMIC KDA MODEL

Massoud Vadoodparast[1], Prof. Abdul Razak Hamdan[1], Dr. Hafiz[1]
[1]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia , 43600, UKM Bangi, Selangor, Malaysia
.

Abstract **– Clustering analysis and Datamining methodologies were applied to the problem of identifying illegal and fraud transactions. The researchers independently developed model and software using data provided by a bank and using Rapidminer modeling tool.**

The research objectives are to propose dynamic model and mechanism to cover fraud detection system limitations. KDA model as proposed model can detect 68.75% of fraudulent transactions with online dynamic modeling and 81.25% in offline mode and the Fraud Detection System & Decision Support System. Software propose a good supporting procedure to detect fraudulent transaction dynamically.

Keywords-component; Fraud detection, Data Mining, Clustering techniques, Decision Support System

I. Introduction

Today's detecting and preventing fraudulent financial transactions especially in credit cards from huge volume of data are playing important role in the banking and financial institutions business. Many researches have used data mining algorithms to detect fraudulent transactions. Normally more than one million transactions are created daily , so detecting process in optimal way is a time consuming process and mostly is done offline in static operation, usually the batch processing is used in specific period like daily, weekly or monthly to discover the fraud. The second issue is the learning machine or supervised algorithm like classification relies on accurate identification of fraudulent and non-fraudulent transactions, however these information usually do not exists or limited. Also, it means preventing of happening fraudulent transaction do not occur in transaction time or the system using predefined rules and scenarios or static model. In order to fill this gap and needs of periodically update of rules to perform optimally, it is necessary to present dynamic models. Thus, the research objectives are to propose dynamic model and mechanism to cover these two issues. The standard data mining methodology is adopted in this research. Table 2 shows the researches have done by researcher based their country; we can see that United State has most part, based on Table 1, we show that the United State suffer for fruad problem with overally 42% in last years , it means US has good approch to manage this problem.

Country	Cardholders Affected (Overall)	Cardholders Affected (Last 5 Years)
United States	42%	37%
Mexico	44%	37%
United Arab Emirates	36%	33%
United Kingdom	34%	31%
Brazil	33%	30%
Australia	31%	30%
China	36%	27%
India	37%	27%
Singapore	26%	23%
Italy	24%	22%
South Africa	25%	20%
Canada	25%	19%
France	20%	18%
Indonesia	18%	14%
Sweden	12%	11%
Germany	13%	10%
Netherlands	12%	8%

Table 1.Cardholders Impacted by Fraud by Country [1]

Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Country	Study	Method	Details
USA	Ghosh & Reilly(1994)	Neural networks	FDS (fraud detection system)
	Ezawa & Norton (1996)	Bayesian networks	Telecommunication industry
	Chan et al. (1999)	Algorithms	Suspect behavioral prediction
	Fan et al. (2001)	Decision tree	Inductive decision tree
	Maes et al. (2002)	Bayesian networks & neural networks	Credit card industry, back-propagation of error signals
UK	Bently et al. (2000)	Genetic programming	Logic rules and scoring process
	Wheeler & Aitken(2000)	Combining algorithms	Diagnostic algorithms; diagnostic resolution strategies; probabilistic curve algorithm; best match algorithm; negative selection algorithms; density selection algorithms and approaches
	Bolton & Hand (2002)	Clustering techniques	Peer group analysis and break point analysis
Germany	Aleskerov et al.(1997)	Neural networks	Card-watch
	Brause et al.(1999a)	Data mining techniques & neural networks	Data mining application combined probabilistic and neuro-adaptive approach
Canada	Leonard (1995)	Expert system	Rule-based Expert system for fraud detection (fraud modelling)
Spain	Dorronsoro et al.(1997)	Neural networks	Neural classifier
Korea	Kim & Kim (2002)	Neural classifier	Improving detection efficiency and focusing on bias of raining sample as in skewed distribution. To reduce "mis-etections".
Cyprus	Kokkinaki (1997)	Decision tree	Similarity tree based on decision tree logic
Singapore	Quah & Sriganesh (2007)	Neural networks	Self-Organizing Map (SOM) through real-time fraud detection system
Ukraine	Zaslavsky &Strizkak (2006)	Neural networks	SOM, algorithm for detection of fraudulent operations in payment system

Table 2, summary of studies investigating different techniques in credit card fraud [2]

II. Overall Process of Data Mining for fraud detection

The overall process of fraud detecting using data mining methods cotains following steps as showon in Figure 1.

- Gathering data of domain and related knowledge
- Selecting transactional dataset based on date and time or quantity or combination of both, customer based or customer group based.
- Preprocessing (Remove noise,Handling missing value,Transformation into suitable form for mining.
- Using Data mining technique which is usully searching patterns based on models such as classification,neural network or outlier recognition based on clastering technique.
- Pattern or outlier evaluation to identify representing knowledge
- Send the extraeted information to DSS in order to decide wither it normal or abnormal behavior

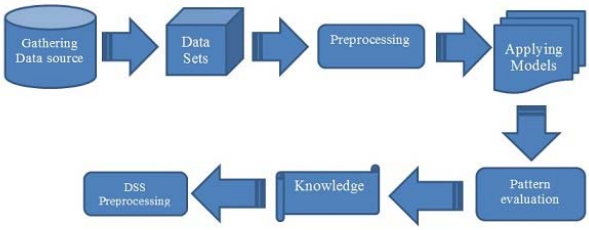


Figure 1. Overall process of fraud detection

Normally these process is done offline or staticly becuase the volume of transaction is huge and this process is time consuming, so providing a dynsmic model for huge volume of data is not esay and processing this model take time ,while the transaction done in less than mili second.

One of the most important challenge is, using supervised data mining technique like learning machine or classification relies on accurate identification of fraudulent and non-fraudulent transactions, however these information usually do not exists or limited or confidential . Financial institutions prefer to not disclose this kind of information and categorize them in high-risk data, so accessing to this kind of data is very restricted. Therefore the process has difficulty in step "Applying Models" and "Pattern evaluation", so the extracted knowledge might not be cover all fraud scenarios and it increase the error and decrease the accuracy and finally the Decision Support System (DSS) accuracy is decreased as well. So many researches have done to fill this gaps and present models or techniques to overcome these issues and enhance the DSS.

III. Data Mining and clustering

In clustering problems usually, we have set of properties or dataset and looking for some similarity or dissimilarity based on some predefine criteria. This similarity criterion case by case is different for different problems. For example if the datasets are contiguous we can use Euclidean distance as similarity criteria[3], so every dataset will map in multidimensional space as point and each dimension represent one feature or property of dataset.

In clustering problems, there is no special class, actually, we do not have class factors as classifier and just based on similarity, the categorization and clustering will be done. The most similar records or dataset will group in same cluster, so the different clusters have less similarity to each other.

Because of we are not defining classifier for clustering algorithm and data do not labeled or tagged, this technique categorize as unsupervised techniques. The clustering results will analysis for extracting order or knowledge from clustered datasets. Clustering outputs

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

reanalysis again in order to find discipline between clusters, the important point is that, always clustering work based on input properties or parameters, same dataset with different parameter might lead to different clustering results and it is not related directly to clustering algorithm.

The aim of clustering is minimize the Intercluster Distance and maximize the Intracluster Distance. (See figure 2 regarding this)

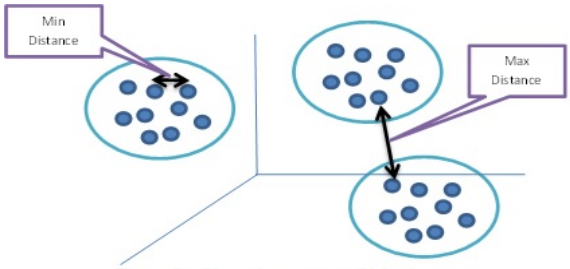


Figure 2. Clustering process Objectives
(adapted from Practical Datamining book,2012)

A good clustering method will produce high quality clusters in which:

- The intra-class (that is, intracluster) similarity is high.
- The inter-class (that is, intercluster) similarity is low.

The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of clustering method measures, by its ability to discover some or all of the hidden patterns as well.

IV. Schematic Overview of Clustering Process in this Paper

In Figure 3 the overall process and steps of fraud detection and DSS are presented.in first step, the historical repository database of previous customer transactions should be prepared and based on model required parameters, the preprocessing are applied. When new transaction comes, based on data window size (that is last 100 transactions, in this paper), customer dataset fetch from repository, and a new transaction is sent to the clustering model, in order to develop a customer behavior model. After applying model, the clustering results will send to a DSS in order to decide whether it suspicious transaction or the behavior is normal. The result evaluation is, based on genuine fraud cases as external dataset evaluation.

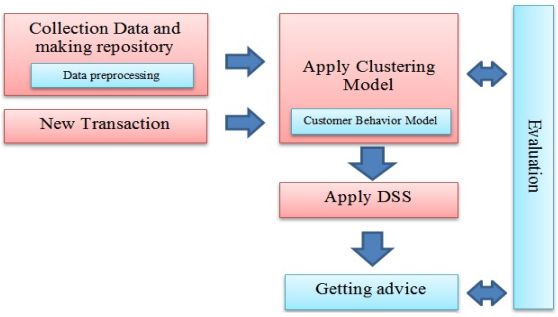


Figure 3. the overall process of FDS in this thesis

In this research, 3,609,618 real banking transaction data for 1015 customers were collected and 32 genuine fraudulent cases are used to compare and to evaluate the result.

The transaction data are preprocessed in order to improve the quality and process speed. The collected data contains 44 items per transaction and we used 8 items for modeling purpose. For the time accuracy the hour accuracy are considered, so to process the historical date so the transaction data should grouped, hour wised.

Data preprocessing step uses for optimizing quality of datasets, in clustering model. However, in this thesis the statistical data is used for all datasets without any elimination because we looking for outlier and abnormality in this model. If we remove outliers, we might lose suspicious transactions. After preprocessing, repository data is ready to use for customer behavior modeling.

In preprocessing data, we filtered the transaction data based on:

- The transaction should be from purchasing type group like retail transaction, bill payment or top up transaction
- The transaction should be settleted
- We extract :
  - PAN for identify the customer
  - TermID for identify the terminal id,normally the customers using same place or same web payment in their transaction.
  - MerchantID to identify the merchant ,normally customers using same merchant for their regullar shopping
  - PosCondition to identify the paymnet device like POS , Mobile,Internet.normally customers have some habit in using media like mobile or POS.
  - AffectiveAmount as transaction amount
  - BusinessDate as transaction date
- We have processed BusinessDate and divided it in two fields: transaction Date and transaction Time based on transaction hour. Normally the customer make their transaction in similar date like end of month and usually in same hour ,especially for bill payment

The result of preprocessing dataset is shown in Table 3 , that will be used in research data mining model as input repository.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

R	Filed Name	Type	Description
1	PrCode	Integer	Process code type of transaction
2	PAN	Varchar	Masked Card NO
3	TermId	Varchar	Terminal identifier
4	MerchantID	Varchar	Merchant identifier
5	PosCondtion	Integer	POS Operation type (Bill,Top up,…)
6	AffectiveAmount	Double	Affective transaction amount
7	TrxDate	Date	Transaction business date
8	TrxTime	Integer	Transaction Hour

Table 3 dataset filed after preprocessing

All algorithms needs one tag as identifer to make it unieque in data set for each record so, we add one more lable as ID to identify each record throug and after processing.

V. KDA Clustering Model

As shown in Figure 4, the final proposed model as KDA clustering model is a combination of three clustering algorithm, K-MEANS, DBSCAN and AGGLOMERATIVE clustering algorithms that represented together as dynamic solution. When new transaction happened, the customer behavior model generate (including new transaction) and the customer dataset cluster with three clustering algorithm, K-MEANS, DBSCAN and AGGLOMERATIVE, it means each record will have three labels that will use to detect abnormality.

Each algorithm might use all or some parameters of prepared dataset. Suspicious transaction will be in the clusters with minimum members in K-MEANS, high LOF values in DBSACN and in a single node in AGGLOMERATIVE algorithm that appear and detected at least by two of clustering algorithms. It means if the new transaction detected by two or more algorithm in as suspicious transaction, it takes place in suspicious area and will potentially fraudulent transaction.

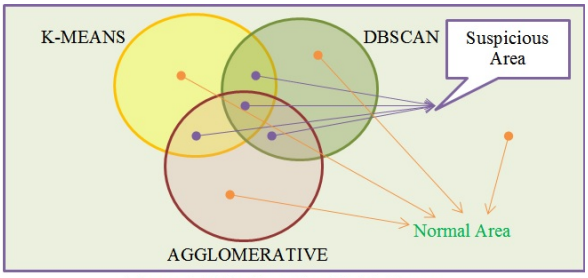


Figure 4. KDA Clustering Model diagram

In this model, the model processing happen parallel for each algorithm and the results will write to separated tables in database, so we can easily with comparing result detect abnormally in customer behavior.

K-MEANS good enough when, we able to define center points and define K as number of clusters and it can detect noise and outlier by measuring distance very good, we can find and optimize center point (here named centroid) by repeating and rerunning the algorithm again on the result of previous execution. So, the problem of this algorithm is finding optimal K.
We can summarize K-MEANS steps as[3]:

- Input : K, number of cluster and n, objects dataset
- Output: set of K cluster with minimum squared errors criteria

Below are algorithm steps:

1) Pick a number (K) of cluster centers - centroids (at random)
2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
   - $d_E(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$
   - $d_H(x,y) = \sum_{k=1}^{n} |x_k - y_k|$

   That, n is number of dimensions or number of dataset properties and x_k and y_k are k properties of x and y objects
3) Move each cluster centre to the mean of its assigned items
4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

In the DBSACN algorithm, the number of the clusters not fixed or predefined, this algorithm looking for point with maximum density in their surrounding neighborhood and number of cluster specified dynamically. The one cluster based on density contains a set of objects that all Density-Connected to each other. That means any object outside these cluster consider outlier or noise.For detecting local outlier, a degree to each object will assign to be an outlier. This degree called the Local Outlier Factor (LOF) of an object. The degree depends on how the object is isolated with respect to the surrounding neighborhood Defining ε as surrounding neighborhood radius is very important in this algorithm because if assign small number, number of the clusters will increase and all data going to separated clusters and if assign big number all data going to one big The clusters will increase and all data going to separated cluster , so finding optimal ε is very important.
We can summarize DBSCAN steps as [3]:

1) Computing (k- distance of p)
2) Finding (k-distance neighborhood of p)
3) Computing(reachability distance, p wrt object o)
4) Computing (the local reachability density of p)
5) Calculation Local outlier factor of p

AGGLOMERATIVE algorithm works like tree, first consider each object as one cluster and then start combining these clusters together based on some criteria and make bigger cluster until all cluster combine and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

make big tree or meeting stop condition[4]. This algorithm works by comparing distance between all objects in same cluster together and divides the objects with maximum similarity in one cluster, and repeat processing with new cluster. In this algorithm if repeat cycle many times, all objects will be take place in one cluster separately and if we run it enough might the results not good to make decision regarding results. Taking place in one cluster is not a problem because the model represent tree and by analyzing tree we can take decision but it time consuming process and might run cycle hundred or more times, therefore stop condition is main issue of this algorithm.

Algorithm	Main Issue
K-MEANS	Assigning proper K
DBSCAN	Defining proper ε
AGGLOMERATIVE	Stop condition

Table 4. Compare Clustering algorithms main issues

In proposed technique, the final decision make based on comparing of output of all algorithms together in order to decrease the errors and increase the accuracy K-MEANS is fast and the accuracy is good but it is static clustering, so we cover it with DBSCAN and AGGLOMERATIVE with dynamic cluttering. DBSACN is dynamic but if fraud happen out of ε radius cannot detect it, but K-MEANS and AGGLOMERATIVE able to detect noise in all distances. AGGLOMEARTIVE is dynamic but not enough fast and might put all object in one cluster specially when increase the parameters, but K-MEANS and DBSACN have stop condition. So, we can conclude these using algorithms together can cover each other to solve the fraud problem better.

VI. Model Specification

For bulding customer behavioral model we select bellow items as K-MEANS dimensions, that means, this model has 6 dimensions.

- AffectiveAmount
- MerchantID
- PosCondtion
- PrCode
- TRXDate
- trxtime

Parameters:

- Number of attribute =6
- K=12
- Max runs=10
- Measure Type= Numerical Measure
- Numerical Measure = Euclidean Distance
- Max optimization step=100

in this model , it is set K=12 that is for last 3 months equal to 12 weeks, the purpose is cluster every week in one cluster if everything be normal, and n=100, that maximum number of transaction in last 3 months. In evaluation phase, we will evaluate the accuracy of K with Davies-Bouldin index calculation as performance evaluation for K-MEANS clustering and prove that the K=12 is the optimal.

For DBSCAN like K-MEANS, 6 items parameters are used and numerical measure as measure type, with Euclidean distance calculation to calculate dependency for detecting noise and outliers are used.
These 6 dimensions for this clustering include:

- AffectiveAmount
- MerchantID
- PosCondtion
- PrCode
- TRXDate
- trxtime

Parameters:

- Number of attribute =6
- Epsilon(ε)=1000000
- Min points=1
- Measure Type=Numerical Measure
- Numerical Measure = Euclidean Distance

The minimum cluster object is set to 1, it means at least the output has one cluster, and $\varepsilon = 1000000$ that the minimum amount that important in banking system to inspect for fraud in fraud detection process (the currency is Rails and this amount equal to 100 Malaysian ringgits)

For AGGLOMERATIVE algorithm, we choose three dimensions for this clustering:

- AffectiveAmount
- TRXDate
- trxtime

Parameters:

- Number of attribute =3
- Mode=Average Link
- Measure Type=Numerical Measure
- Numerical Measure = Euclidean Distance

To simplify the complexity of this algorithm, the parameters are reduced to three fields, Average Link are recruited. Numerical measure as measure type with Euclidean distance is selected as well.

These three clustering algorithms works with numerical data not nominal, so converting data to numeric is one pre step before running the model. We used convertor adapter to convert nominal data to numeric, and all data converted to numeric.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

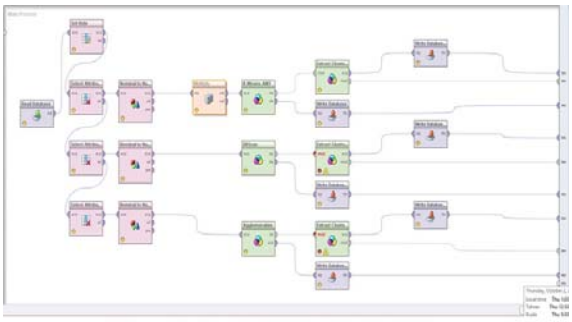


Figure 5. Implemented KDA model in RapidMiner

Figure 5 shows the implemented KDA dynamic model in RapidMiner software.

When new transaction happened, the customer behavior models generate for these three selected algorithms (including new transaction) and suspicious transaction will take place in shared space between at least two algorithms that usually, are in the clusters with minimum members and high LOF values or in a single node. So ,the KDA model space is shared spaces between these three algorithms that each algorithm try to detect abnormality with different technique, on the other hand, overlapping areas are as desire area and required answer for fraud detection problem.

In this model, the model processing happened parallel for each algorithm, it means, we are checking distance, density and objects route link together in same time and then deciding regard occurred transaction, we try to see transaction from different perspective to make sure detecting process work optimally. The results of each algorithm write to separated tables in database, so we can easily detect abnormally in behavior with comparing result.

VII. Discussing FDS&DSS Logic

Decision support system regarding fraud detection is a one of most important section in all financial organization, that wrong decision influence directly the business and it causes dissatisfaction in customer area. Therefore, the decision rules and policies are normally conservative and somehow managers prefer to inspect issues manually or just getting advices form Fraud Detection System (FDS) regarding stop online suspicious transaction specially when new scenarios happening. With growing fraudulent transactions in last years the approach of using automated FDS is increased and many FDS are developed.

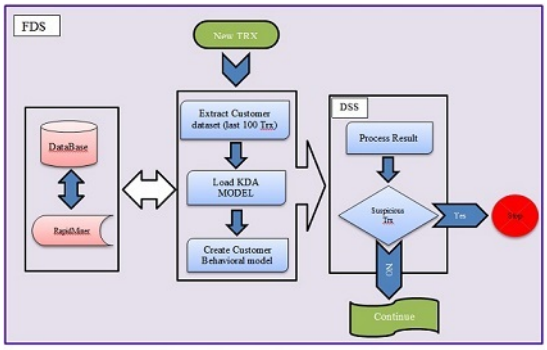


Figure 6. Proposed FDS diagram

In proposed DDS that use KDA model for detection fraud inspect the suspicious area and if transaction take place in this area the system will arise alert to advise user to inspect the transaction or stop it.

The DSS logic is simple and works as:

If trx(n) detected by K-MEANS(n) as Fraud then
nK=1
Else
nK=0
If trx(n) detected by DBSCAN(n) as Fraud then
nD=1
Else
nD=0
If trx(n) detected by AGGLOMERATIVE(n) as Fraud then
nA=1
Else
nA=0
If (nK and nD) or (nK and nA) or (nD and nA) then
nF=1 ;
SendAlert;
Else
nF=0
Continue;

Based on bank policy, the system can stop the suspicious transaction or just raise alert for user in order to inspect the transaction.

VIII. Discussion & Results

As definition ,we have defined :

- True Positive Rate (TPR) → Normal transactions and model detect normal
- False Positive Rate (FPR)→ Abnormal transactions and model detect normal
- True Negetive Rate (TNR) → Normal transactions and model detect Abnormal
- False Negetive Rate (FNR)→ Abnormal transactions and model detect Abnormal

This model aimed to increse True Positive Rate (TPR) and False Negetive Rate (FNR), it means increase accuracy regarding normal and abnormal transactions, and decrease False Positive Rate (FPR) and True Negetive Rate (TNR) means reducing errors. On the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

other hands, the system detect normal and abnormal transactions properly and reduce errors in this process.

we have run this model for 100 customer that already have normal transactions in Databases and investigate the results.

R	Model	TPR	TNR
1	K-MEANS	90	10
2	DBSCAN	84	16
3	AGGLOMERATIVE	88	12
4	KDA Model	96	4

Table 5 ,Model results for Normal transactions

Results show , the KDA model can detect 96% of normal transaction properly. Logic of KDA is based on, if two model detect one transaction as normal transaction , the result will be normal and this optimazation is becuse of using more than one clustering technieque in the final model. We can see, if we using one clustering model, in best state the result will be 90% that related to K-MEANS, but here the KDA model accuracy is 96% , it means atleast 6% of normal transactions in K-MEANS detect as abnormal, on the other hand the KDA model optimize error of K-MEANS 6%, DBSCAN 12% and AGGLOMERATIVE 8% as well.

From other point of view, we can see, at least 6 transactions exist that K-MEANS algorithm cannot detect it properly but DBSCAN and AGGLOMERATIVE can detect them better. Figures 7-10 show model output from RapidMiner software.

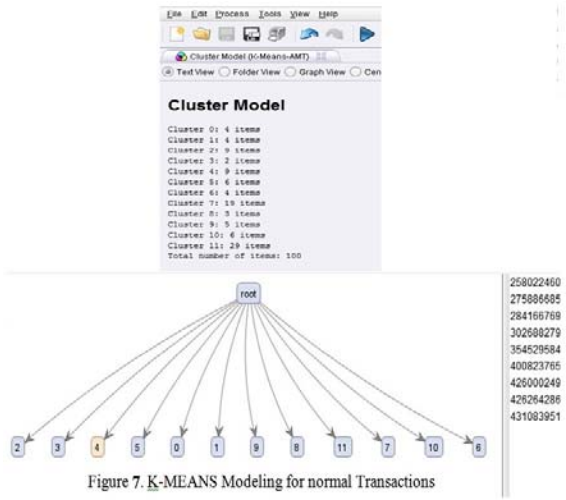


Figure 7. K-MEANS Modeling for normal Transactions

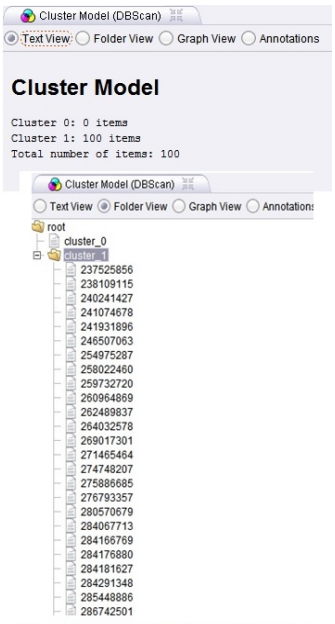


Figure 8. DBSCAN Modeling for normal Transactions

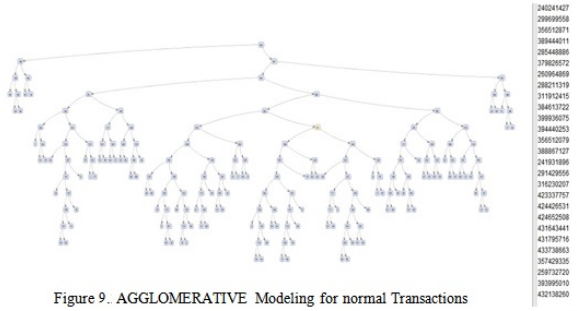


Figure 9. AGGLOMERATIVE Modeling for normal Transactions

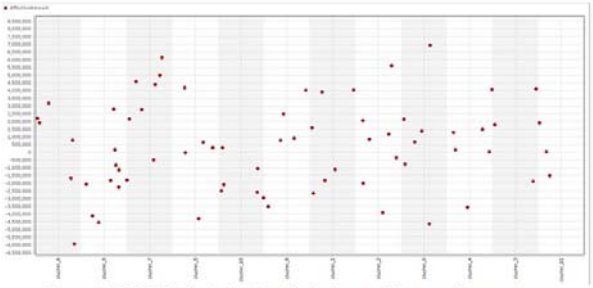


Figure 10. K-MEANS clustering for AffectiveAmount for normal Transactions

For testing the model with genuine cases as external evaluation ,we have run this model for 32 fraudulnet transaction,as mentioned previously ,the Database has 1015 customers information.

In first step , we have run the model for all historical customers data , in this period all fraudulent transaction is 16 and the model has detected 18 transaction as fraud , from this 18, 13 was correct , it means FNR=13 and TNR=5 and model could not detect 3 transations at all and detect them as normal transations, it means FPR=3.Table 5.3 shows the KDA model resultant is better then each model seperatly, we can see K-MEANS model is more sensetive thant two other models but the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

precision is lower (FPR, TNR is bigger) and agglomertive detection is less sensetive but false detection is better (FPR, TNR is smaller).Results are shown in Table 6.

R	Model	Total Detect	TNR	FNR	FPR
1	K-MEANS	21	10→62.5%	11→68.75%	5→31.25%
2	DBSCAN	19	7→43.75%	12→75%	4→25%
3	AGGLOMERATIVE	17	6→37.5%	11→68.75%	5→31.25%
4	KDA Model	18	5→31.25 %	13→81.25%	3→18.75%

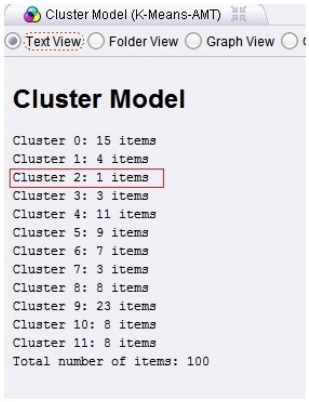
Table 6.Model results for fraud detection

In next step , we test the model with genuine fraudulent cases in real time to see the result of dynamic modeling. We test the model with 16 frudulent transactions. The result are shown at Table 7

R	Model	FNR	FPR
1	K-MEANS	9 →56.25 %	7→43.75 %
2	DBSCAN	7 →43.75%	9→56.25%
3	AGGLOMERATIVE	8 →50%	8→50%
4	KDA Model	11→68.75%	5→31.25

Table 7.Model results for real time fraud detection

The results show, the KDA model still have better results than each model separately, with combination of each model results with this logic: “if the transaction detected by two model as fraud so in KDA model consider as suspicious transaction”, the final results is 68.75 %.of fraudlent transactions can be detected by this model.Figures 11-15 show model output from RapidMiner software.



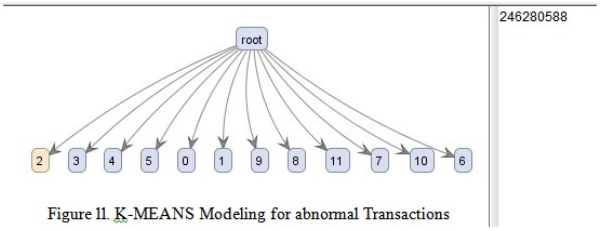


Figure 11. K-MEANS Modeling for abnormal Transactions

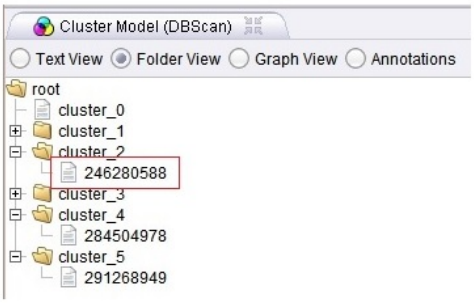


Figure 12 . DBSCAN Modeling for abnormal Transactions

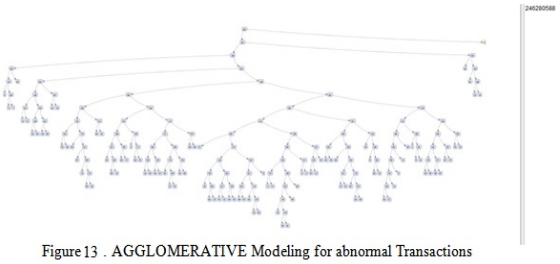


Figure 13 . AGGLOMERATIVE Modeling for abnormal Transactions

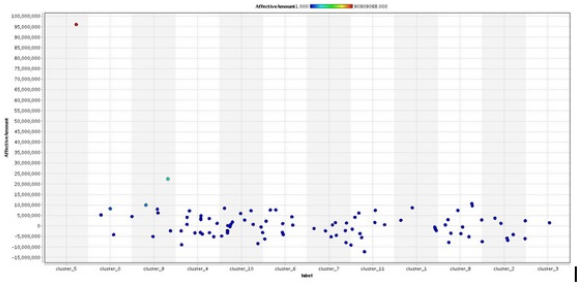


Figure 14 . K-MEANS clustering for AffectiveAmount for abnormal Transactions

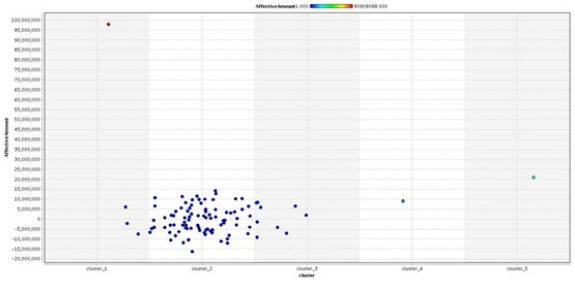


Figure 15. DBSCAN clustering for AffectiveAmount for abnormal Transactions

IX. FDS & DSS Outputs

Developing FDS and DSS as software in order to help and advice inspector to inspect transaction faster with more accuracy is the last part of this research. When software load RapidMiner KDA model, model and its objects load in the memory and can interact directly with software and database as well. FDS has developed with

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Viusal Vb.Net 2010 and the Database Engine is Microsoft Sqlserver 2008.

In the DSS , two options provided in the software, first process the offline transactions, it mean we can run the system and check previous customer transactions by clicking on “Process Historical Data” button and second option process new transaction. For simulation purpose, we can add new transaction manually and process it. Definitely, in the online mode, the database updated automatically so no need to use this option. Figures 16-21show some FDS software outputs.

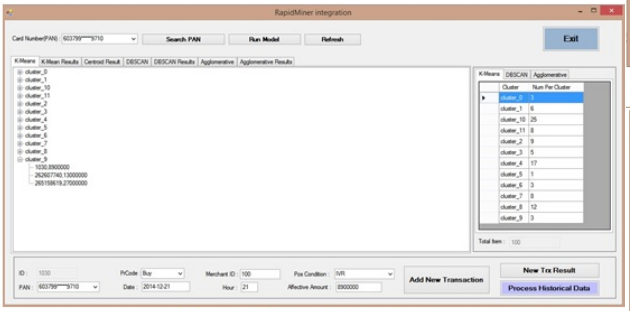


Figure 16. K-MEANS Tree Result

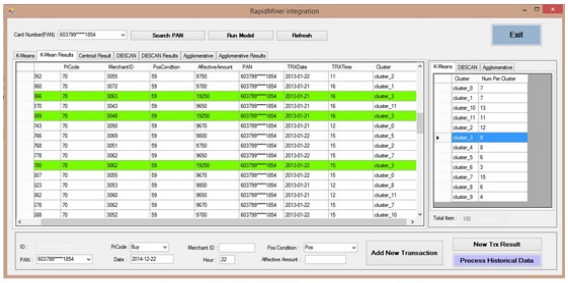


Figure 17 . highlighting cluster members

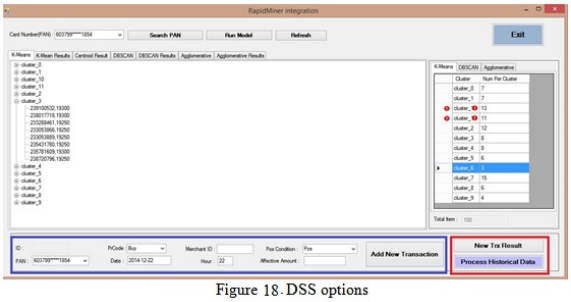


Figure 18. DSS options

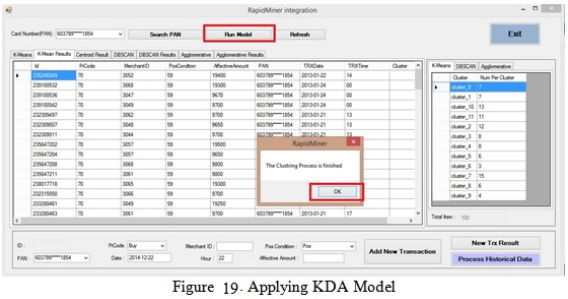


Figure 19. Applying KDA Model

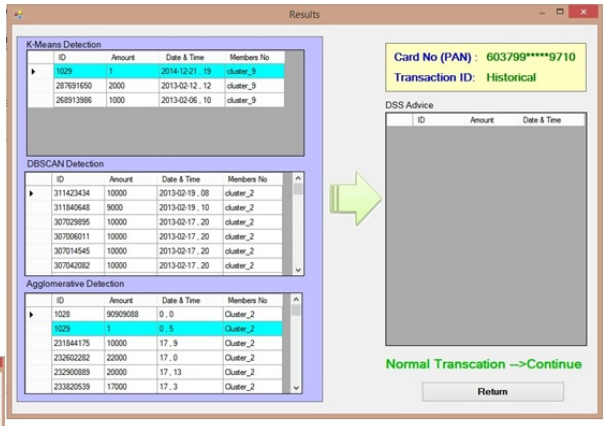


Figure 20. DSS result for normal historical dataset

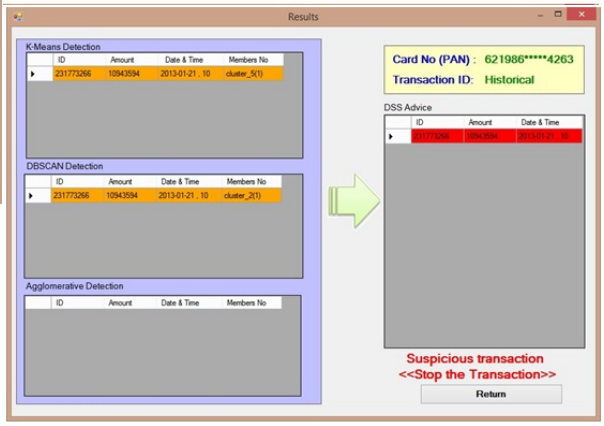


Figure 21. DSS result for abnormal historical dataset

X. Conclusion

KDA model could improve consuming time processing and make three customer modeling in the same time to help detection suspious transaction in customer side better. Devloped FDS and DSS softwares can highlight and then classify the transaction with result of modeling.

The accuracy obtained by KDA modeling is 68.75% for dynamic online modeling and 81.25 % for historical or offline modeling and seemed it is competitive with other algorithms in this area.

References

[1]John Kiernan,2013 ,Credit Card & Debit Card Fraud Statistics,http://www.cardhub.com/edu/credit-debit-card-fraud-statistics/(2013.)

[2] Linda Delamaire , Hussein Abdou , John Pointon (2009). Credit card fraud and detection techniques , Banks and Bank Systems, Volume 4,57-68

[3] M Saniei A,S Mhmoudi & M Taherparvar, (2012),Practical Data Mining,Nayaz Danesh Publication, ISBN :978-600-6481-12-8

[4]M Ghazanfari,S Alizadeh & B Taymorpour (2011),Data Mining & Knowledge Discovery,IUST Publication, ISBN:964-454-178-2

[5] Fareed Akthar, Caroline Hahne,(2012), RapidMiner 5 Operator Reference
http://www.RapidMiner.com /(24th August 2012)

Comparison between C++ console and graphic programming

Zhengyu Lu
Department of Computer Science
Jilin University
Jilin Province, China

Dimitar Pilev
Department of Informatics
University of Chemical Technology and Metallurgy
Sofia, Bulgaria

Abstract: Nowadays, the new students of Computer Science learn programming starting with C or C++ in the black-white console. At first, they may feel very proud of themselves by calculating or print the result on the Screen. As time goes by, they may feel tired of the black-white box and prefer something new. Then the problem comes, they should learn the basic C++ code by heart or just skip it to the graphic programming Which sometimes seems more interesting than black-white console programming. And after the Graphic programming like MFC, CLR programming, we can even release the program and make it a small software that can be set up in their own laptop. It is really more interesting. So we will look at one example that we build the program through console and released-version graphic programming to see the difference and disadvantages and advantages of them.

Keywords-C++; Console programming; Graphic programming; Comparison;

1. Introduction of a program

In our daily life, sometimes we are required to predict some unknown elements from what we have already known. For instance, some scores of a number of students' subjects are known,

while the rest of them are unknown and then we can create an application to calculate the unknown ones. Now we are going to show a paper work in which we can create such a application. In this C++ program, there are several requirements that we need to follow and formulas to use accordingly. Here are some details from the project.
In this paper work, we are going to deal with a matrix with 6 rows and 12 columns which stand for 6 students and 12 degrees of subjects from 1 to 5. 0 represents the ones we do not know. For example, the score of the first subject of the first student is 1, however the score of the second subject is unknown. Then we are going to create one application that can solve such a problem.

1 0 3 0 0 5 0 0 5 0 4 0
0 0 5 4 0 0 4 0 0 2 1 3
2 4 0 1 2 0 3 0 4 3 5 0
0 2 4 0 5 0 0 4 0 0 2 0
0 0 4 3 4 2 0 0 0 0 2 5
1 0 3 0 3 0 0 2 0 0 4 0

Encountering such a project, students learning C++ may choose different programming tools. In

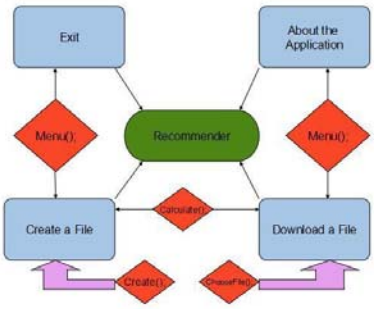
http://sites.google.com/site/ijcsis/
ISSN 1947-5500

this essay, we based on Microsoft Visual Studio 2010 to solve it using console and windows programming. In the end, we compare the advantages and disadvantages of them.

2. Console Programming

We are going to create a menu with four choices for the user to choose. They are **“ Exit ”**, **“ About the Application ”**,**” Create a File”** and **“ Download a File”**. In the **“ Exit”**, the user will kill the application. In **“ About the Application ”**, we will provide the user with the chief information of the application. And in **“ Create a File”** and **“ Download a File”**, the user can create or download a file from the disk.

Now the next is to create different functions to fulfill the requirements of the four buttons.



As we have shown, the application named **“ Recommender”** have four branches and each of them are made up of different functions. We will combine the running status and functions together to show how the application works.

Here is the first running step of the application.

In the function of **menu()**, we set a loop and it will not stop until the user input **“1”** to choose **“ Exit”**.

```
void menu(){
                bool check=true;
                while(check){
                        welcome();
        ...
        }
```

In the choice of **“ About us”**, we can print the general information of the application easily.And if the user choose **“ 3. Load a file”**, the application will provide the user with several

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

choices first and read the file from disk after the
user make a choice. After printing the
information read from disk, it is time to predict the unknown ones now. Predictions can be calculated with the following formulas.

$$r_{ui} = R_u + K_{ui} \sum_{v \in U_i(u)} w_{uv}(R_{vi} - R_v)$$

$$w_{uv} = \frac{\sum_{j \in I_{uv}} (R_{uj} - R_u)(R_{vj} - R_v)}{\sqrt{\sum_{j \in I_{uv}} (R_{uj} - R_u)^2 \sum_{j \in I_{uv}} (R_{vj} - R_v)^2}}$$

$$K_{ui} = 1 \Big/ \sum_{v \in U_i(u)} |w_{uv}|$$

where:
Ru is the average of the ratings given by a user u (i.e., the average of the numbers Rui such that Rui !=0).
Iuv is the set of items that both u and v have rated, i.e., the set of indexes i such that Rui!=0 and Rvi!=0.
wuv is the Pearson correlation of the users u and v. Note that wuv is not defined and cannot be computed if Iuv does not contain at least 2 items. Moreover, note that the averages Ru and Rv are here computed considering only on the items that are in Iuv.
Ui (u) is called a neighbor set of u, and it is the set of users v that have rated item i, i.e., all the indexes v such that Rvi!=0 and for which you can compute wuv (hence v and u must have rated at least 2 common items).
Kui is the reciprocal of the sum of all the absolute values of the Pearson correlations of u with all the other users that have rated i.

In the function of **calculate()**,

```
void calculate(int *p,int u,int i){
...
}
```

We provide three parameters named p, u and i. P represents the array used for storing the elements read from file. U represents the number of student(user) and i represents the number of(item). Firstly, we should get all the parameters used in the formulas and it is easy but just calculations. Secondly, put all the parameters into formulas and predict the unknown ones.

In the function of **create()** without parameters,
Firstly, we are going to store the ratings user input in array and then change them into char type in order to write them in file.

```
Ofstream write("D:\\Project\\Created.txt");
        int *integer=new int[u*i+1];
        char * str=new char[u*i];
```

Secondly, write the original file into disk.

```
if(write.fail()){
        cout<<"Open error !"<<endl;
        exit(1);
}
```

Next, predict the unknown ones and write them back in files also.

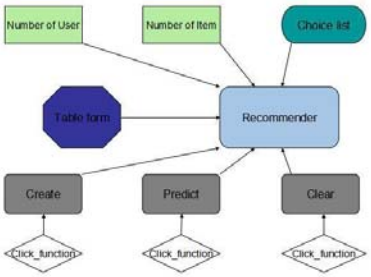
```
write.open("D:\\Project\\updateCreated.txt");
        write<<"    Welcome and this is your updated file"<<endl;
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Up to now, we can get the results of unknown numbers.

Windows Graphic Programming

We follow the same examples and formulas to show how it works. For example, for the prediction part, we have the same test data and predicting formulas. In a word, we set two different labels which shows the number of students(users) and subjects(items), one choice list that provide users with various files to download, one table form for user to create and show the prediction results on screen and three different buttons acting as **"create"**, **"predict"** and **" clear".**



About the labels, here we give an example of the code to show how it works. This is **"label4"** which refers to **"Number of Users"**.

```
this->label4->AutoSize = true;
this->label4->Location =
System::Drawing::Point(31, 26);
this->label4->Name = L"label4";
this->label4->Size = System::Drawing::Size(95, 12);
this->label4->TabIndex = 11;
this->label4->Text = L"number of users";
```

Through these properties, we can control the position, size, name and so on of the label.

In the **choice list,** we also provide two choices for user and then download the elements from file to the form. Here is the code for choice list.

```
if(curItem=="text.1"){                                read.open("D:\\Project\\test1.txt"); //do not
forget to use "using namespace std;"
                                }
else{
        if(curItem=="text.2"){
                read.open("D:\\Project\\test2.txt");
                                        }
        else{
        read.open("D:\\Project\\test3.txt");
           }
  }
```

In the **create button**, the application can create a table form for user according to his or her mind. Here is the example of creating columns in the form.

```
for(int i=0;
i<this->dataGridView1->RowCount; i++){    // Create the first column,
        if(i<rowNum){                     // According to the number the user input.
        dataGridView1->Rows[i]->Cells[0]->Value=System::Convert::ToString(i+1);
                                }
     else{                                dataGridView1->Rows[i]->Cells[0]->Value="";
```

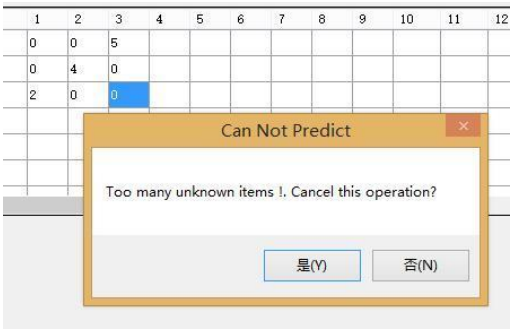

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
            }
    }
```

Next in the **clear button**, we set the value of every element to blank. Here is the original code of how it works.

```
for(int i=0;i<this->dataGridView1->RowCount;i++){
      for(int j=1;j<this->dataGridView1->ColumnCount;j++){
          dataGridView1->Rows[i]->Cells[j]->Value ="";      // Set every element in the table into blank
                    }
          }
```

Next in the **predict button**, the same like in console programming, we use various functions to predict the unknown numbers. What is more, we provide the function of reminding the users of errors which may happen because of too many unknown numbers.



Last but not least, we can release this application which ca be set up and used in more computers.

Comparison between **console programming** and **graphic programming.**

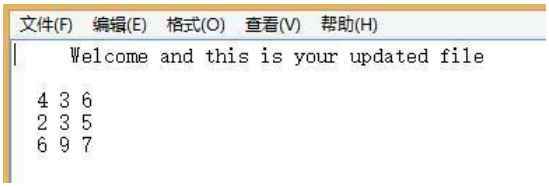
Differences

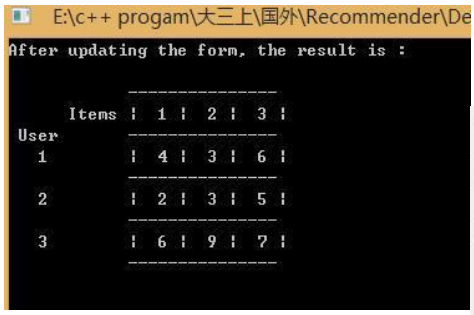
First, it is clear that we can only control the application by keyboard but not by mouse which is

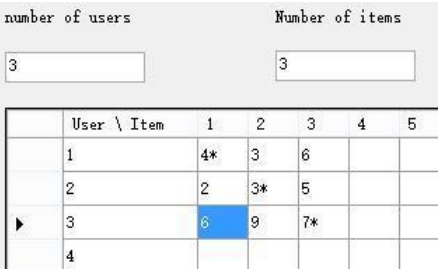
http://sites.google.com/site/ijcsis/
ISSN 1947-5500

created by console programming. But we can click or press mouse to control the application created by graphic programming, which is more convenient. And also, with so many buttons and labels, it is more vivid to control the application created by graphic programming.

Secondly, the programmer has to create a form with elements by himself or herself, which sometime is hard to control the rows and columns. In this example, we prefer to output the new elements with asterisk, which works better in graphic programming than in console programming.



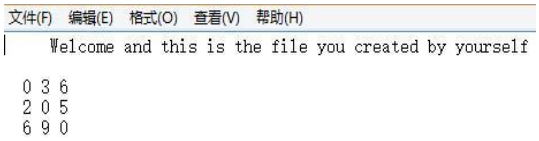




Another thing, when we are using console programming, we can print the numbers no matter known or unknown directly in the screen, but for graphic programming it is different. We need to convert the numbers from float type to char type, so that they can be printed in the form, which may be a little complicated when compared with console programming.

```
for(int i=0;i<rowNum;i++){     //To get the numbers in the table.
        for(int j=0;j<columnNum;j++){
        data[i][j]=System::Convert::ToInt32(dataGridView1->Rows[i]->Cells[j+1]->Value);
                                                                    }
                                                            }
```

Next one, when we provide the functions to create and update files in console programming, which is more multi-functional. For user, he or she can store the files in disk and check it in the future. But for many factors, we do not provide in graphic programming.



Another one, about the size of form that we can use to store the numbers. In console programming, we can not change the size of the window so that we can only show a limited number of elements to users. While in graphic programming, it is very different and we can input much more elements in the table. But here, we only provide 20 rows and columns as an example. And also in graphic programming, users can change or correct the numbers they have inputted whenever they want before predict, which dose not work so well in console programming.

Last but not least, about the choice list. In console programming, we can only provide few choices for users because of the limited size of window. While in graphic programming, the user have as many

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

choices as they can, which benefits from the choice list that does not appear in console programming.

Similarities

Both programs have two the same functions which are responsible for **" Predict" and " Create"**.
And also these two functions use the same way to predict and share similar formulas to complete the calculations.

Another one, we use the same three files to test the accuracy of two applications. And of course, the predicting results are the same.

Both of them can be set up in users' computer with setup project created in the solution.

5. Conclusion

Advantages and Disadvantages

Console Programming

In console programming, programmer has to create the form by himself which takes a longer time than graphic programming.

Moreover, when programmer is creating the table, there are many factors to consider, such as how many digits the number has and how many lines and columns the form needs, which is hard to print accurately.

The advantage is console program is smaller than graphic program which takes more codes.

Graphic Programming

In graphic programming, the programmer can design the form by himself and what is more, visual studio 2010 can create these different buttons, labels and choice lists automatically, Which is very convenient for programmer.

Obviously, the application created by graphic programming is more vivid and has a better interaction with the users. Also it is much easier to use the application with clicking mouse to control it.

The disadvantage is with better appearance, the application needs more codes than console programming and as a result it takes more space to install in your computer.

For C++ beginners

Though graphic programming has better appearance and interaction with users, but we

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Suggest that beginner users of c++ should acquire the basic programming skills and knowledge first. After a long time practice of console programming, it will be easier and better for them to learn to use graphic programming.

AUTHORS PROFILE

1. Zhengyu Lu --- Junior student of Computer Science and Technology in Jilin University, China. From Sep. 30, 2014 to Mar 19, 2015 studied in Chemical Technology and Metallurgy, Sofia, Bulgaria as an Erasmus student. During this time, learned and cooperated with Dr. Pilev of Informatics Department in UCTM.

2. Dimitar Pilev --- Doctor of Informatics in Chemical Technology and Metallurgy in Sofia. Taught and guided Zhengyu Lu in C++ program work.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS AUTHORS' & REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2015
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/