IJCSIS Vol. 13 No. 4, April 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

CiteSeer^x beta

dblp.uni-trier.de
Computer Science
Bibliography

Q·Sensei BETA

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 Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.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 and security. 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 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: 555, No. of Citations: 1305)***. Abstracting/indexing/reviewing process, editorial board and other important information are available online on homepage. 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. 4, April 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

Paul D. Nugent, Ph.D., Emilio Collar, Jr., Ph.D. Management Information Systems, Ancell School of Business Western Connecticut State University Westside Classroom Building, Room 203 181 White Street, Danbury, CT 06810, USA

Abstract — Given the current state of cyber crime, cyber attacks, and cyber warfare, it is easy to argue that steps taken by those in cybersecurity and Information Assurance (IA) roles to thwart these attacks deserve heroic status. Yet, the reality is that these functions are rarely perceived by their organizations or by broader society as heroic and this has negative consequences for job satisfaction and for the attractiveness of careers in these fields. This paper explores the concept of the hero broadly as well as in organizational literature to understand how heroes are made in organizations and why the nature of cybersecurity and IA work present barriers to perceptions of heroism. This is because the intrinsic nature of security work focuses on vulnerabilities and therefore differs from other types of work that are focused on new capabilities. The paper concludes with practical recommendations on how managers, industry leaders, and educators can take steps to overcome these limitations and make careers in cybersecurity and IA more attractive.

Keywords: Heroes; Heroism; Cybersecurity; Information Assurance; Information Security

Akhila G. S, Department of Computer Science and Engineering, Mohandas College of Engineering, Anad, Trivandrum, Kerala, India
Mr. Prasanth R.S, Department of Computer Science and Engineering, Mohandas College of Engineering, Anad, Trivandrum, Kerala, India

Abstract — Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. This paper proposes a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Present an algorithm, namely GreedyIL, for runtime generalization. The experimental results show that GreedyIL performs efficiently.

Keywords—Personalized Web Search,User profile

Ehsan Azimirad, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran
Javad Haddadnia, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran

Abstract — In this paper, a precise description of the threat evaluation process is presented. This is followed by a review describing which parameters that have been suggested for threat evaluation in an air surveillance context throughout the literature. Threat evaluation is a critical component of the system protecting the defended assets against the hostile targets like aircrafts, missiles, helicopters etc. The degree of threat is evaluated for all possible hostile targets on basis of heterogeneous parameter values extracted from various sensors, to improve the situational awareness and decision making. Taking into consideration the amount of uncertainty involved in the process of threat evaluation for dynamic targets, the fuzzy logic turns out to be a good candidate to model this problem. This model is based on a Fuzzy logic approach, making it possible to handle imperfect observations. The structure of the

Fuzzy Logic is described in detail. Finally, an analysis of the system's performance as applied to a synthetic static scenario is presented. The simulation results are acceptable and fine and show that this model is reliability.

Keywords-component; Threat Assessment; Fuzzy Knowledge Based System; Decision Support System; Weapons Assignment; Threat Evaluation Fuzzy Model

4. Paper 31031518: A New Data Fusion Instrument for Threat Evaluation Using of Fuzzy Sets Theory (pp. 19-32)

Ehsan Azimirad, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran
Javad Haddadnia, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran

Abstract — This paper represents an intelligent description of the threat evaluation process in 3 level of JDL model using of fuzzy sets theory. The degree of threat is evaluated for all possible targets to improve the situational awareness and decision making in command and control system and is calculated to precise weapon assignment. Taking into consideration the amount of uncertainty involved in the process of threat evaluation for dynamic targets, the fuzzy set theory turns out to be a good candidate to model this problem. In this approach, based on a fuzzy logic, is making it possible to handle imperfect observations. The structure of the fuzzy expert system based on fuzzy number approach is described in detail. Finally, an analysis of the system's performance as applied to multiple dynamic scenarios is presented. The simulation results show the correctness, accuracy, reliability and minimum errors in the system is designed.

Keywords-component; Fuzzy Number, JDL Model, Decision Support System, Dynamic Air Targets, Multi Sensor Data Fusion; Weapons Assignment

5. Paper 31031526: Optimizing TCP Vegas for Optical Networks: a Fuzzy Logic Approach (pp. 33-45)

Reza Poorzare, Department of Computer Science Young Researchers Club, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Shahram Jamali, Department of Computer Engineering University of Mohaghegh Ardabili, Ardebil, Iran

Abstract — Performance of TCP is reduced over buffer-less optical burst switched (OBS) networks by misunderstanding of the congestion status in the network. In other words, when a burst drop occurs in the network and we cannot distinguish congestion and burst contention in the network, TCP wrongly decreases the congestion window size (cwnd) and causes significant reduction of the network performance. This paper employs the fuzzy logic to solve this problem. By using the fuzzy logic we provide a framework to distinguish whether the burst drop is due to the congestion or is due to the burst contention. The full approach, for detecting state of network, relies on Round-Trip-Time (RTT) measurement only. So, this is an end-to-end scheme which only end nodes are needed to cooperate. Extensive simulative studies show that the proposed algorithm outperforms other TCP flavors such as TCP Vegas, TCP Sack and TCP Reno, in terms of throughput, packet delivery count and fairness.

Keywords — Fuzzy Logic, Optical Burst Switching, TCP Vegas, Transport Control Protocol (TCP).

6. Paper 31031532: Improved Algorithm for fusion of Satellite Images Using Combined DWT-FDCT Transforms (pp. 46-50)

Manjushree B S, CSE, DBIT/VTU, Bangalore, India
Shruthi G, CSE, DBIT/VTU, Bangalore, India

Abstract - Image fusion based on the Fourier and wavelet transform methods retain rich multispectral details but less spatial details from source images. Wavelets perform well only at linear features but not at non linear discontinuities because they do not use the geometric properties of structures. Curvelet transforms overcome such difficulties in feature representation. A novel fusion rule via high pass modulation using Local Magnitude Ratio (LMR) in Fast Discrete Curvelet Transforms (FDCT) domain and Discrete wavelet transforms (DWT) is defined. For experimental study of this method Indian Remote Sensing (IRS) Geo satellite images are used for Pan and MS images. This fusion rule generates HR multispectral image at high spatial resolution. This method is quantitatively compared with Wavelet, Principal component analysis (PCA) fusion methods. Proposed method spatially outperforms the other methods and retains rich multispectral details.

Keywords: Image Fusion, Fast Discrete Curvelet Transforms, Discrete wavelet transforms, Local Magnitude Ratio (LMR)

7. Paper 31031534: Biometric Student Record Management System (pp. 51-62)

Onuiri Ernest E., Oludele Awodele, Oshilagun Ibukun, Yadi Chukwuemeka and Etuk Otobong
Department of Computer Science, Babcock University, Ilishan-Remo, Ogun State; Nigeria

Abstract – Information is an important part of any system. In the academic world, information is especially very important and essential. Students have to register for courses, take attendance, quizzes, and exams and as well as check their scores. Years after graduating from the school, students come back asking for transcript. It is therefore very important to handle students' records in a way that is accessible, maintainable and secure. The manual method of cumulating and storing student record is often prone to various degrees of human error and is also unsecured making it exposed to unauthorized personnel. This paper presents the design and development of a biometric student record management that provides an interface between student and the institution to enable prompt checking of grades, as well as track their progress and efficiently record each student's attendance for every lecture attended through the use of a biometric device. The methodology used in developing this system is the waterfall methodology and this was used because it is a one dimensional model, meaning it is very easy to implement and also the documentation is done at the beginning of the software development. During the course of this research, it was realized that developing a biometric student record management system was a herculean task. This system was given to random students to use and 90% of them loved the interactive nature of the system. A projection of record growth in relation to student population and system requirement was carried out in the study.

Keywords – Fingerprint, Biometrics, Biometric Student Record Management System (BSRMS), Student Information Management System (SIMS), Grade Point Average (GPA), Students

8. Paper 31031538: Classification Framework Based on C4.5 Algorithm for Medicinal Data (pp. 63-67)

Karthik Ganesan, Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, India

Abstract - This study proposes a framework with preprocessing techniques namely Missing value replacement, Discretization, Principal Component Analysis (PCA) to extract the key features and then applying c4.5 classifier algorithm to enhance the classification of medicinal data. The input data gets subjected to missing data imputation through any one of the standard methods like mean, mode, constant and manual input. The dataset is then subjected to Discretization to formalize a reasonable set of discrete bins. PCA is then applied on the dataset to identify the principal components of the dataset, which attribute to the mean data inference. C4.5 algorithm has been used to construct a decision tree based on the information gain of the training set. This work used Cleveland heart disease dataset, obtained from UCI machine learning repository. The dataset is composed of details of about 303 patients and helps to predict presence or absence of cardio vascular disorder based on 75 attributes. The proposed framework was applied on this dataset and exhibited an accuracy of about 77.73%.

Keywords — PCA, Discretization, C4.5, classification

9. Paper 31031539: Energy Efficiency of IEEE 802.15.6 MAC Access Modes for Remote Patient Monitoring Applications (pp. 68-77)

Anas Bouayad, Nour El Houda Chaoui, Mohammed El Ghazi, Molhime El Bekkali

1234 TTI Laboratory, USMBA, FEZ, MOROCCO

Abstract - The progress that has been made over the last decade in the medical field was focused on integrating communication and information technology especially Wireless Body Area Networks (WBANs) in healthcare systems for remote patient monitoring (RPM) applications. WBANs have shown great potential in improving healthcare quality, allowing continuous patient to be remotely monitored and diagnosed by doctors. WBAN operates in close vicinity to, on, or inside a human body and supports a variety of medical applications. Energy consumption is a key WBANs since energy-constrained sensors monitor the vital signs of human beings in healthcare applications. In this work, we are interested in evaluating access methods and 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 decrease the energy consumption. 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, (MAC) protocols, access methods, polling, CSMA/CA, Energy consumption.

10. Paper 31031540: Designing a jitter buffer for QoS improvement in VoIP networks (pp. 78-83)

Negar Chehregosha, Dept. of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
Mohammad Ali Pourmina, Dept. of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract -- Today main challenge in IP networks engineering is simultaneous support of different applications such as sending voice, video and data, with appropriate quality of service. The generated traffic by IP telephone, voice and video conference and on line applications, are real time and time sensitive. Jitter is an usual problem in quality of service of VoIP network. The purpose of this paper is to reduce jitter to improve quality of service. Achieve Real time voice quality is required jitter smoothing in receiver that usually is done by jitter buffer mechanism. Here we introduce an algorithm to design jitter buffer. We simulate one VoIP network by OPNeT simulator and Matlab software is used to implement the algorithm; then we compare simulation results before and after applying the algorithm and the effects of changes in buffer size on delay and jitter are checked. Output voice quality will be measured based on PESQ, according to ITU-T P.862 recommendation. The results show packet buffering reduces packets delay and makes values of them become closer together.

Keywords: VoIP, Jitter, Jitter buffer, Delay, Quality of Service

11. Paper 31031541: Trend Analysis in Academic Journals in Computer Science Using Text Mining (pp. 84-88)

Adebola K. Ojo and Adesesan B. Adeyemo
Department of Computer Science, University of Ibadan, Ibadan, Nigeria

Abstract — Text mining is the process of discovering new, hidden information from texts- structured, semi-structured and unstructured. There are so many benefits, valuable insights, discoveries and useful information that can be derived from unstructured or semi- unstructured data. In this study, text mining techniques were used to identify trends of different topics that exist in the text and how they change over time. Keywords were crawled from the abstracts in Journal of Computer Science and Technology (JCST), one of the ISI indexed journals in the field of Computer Science from 1993 to 2013. Results of our analysis clearly showed a varying trend in the representation of various subfields in a Computer Science journal from decade to decade. It was discovered that the research direction was changing from pure mathematical foundations, Theory of Computation to Applied Computing, Artificial Intelligence in form of Robotics and Embedded Systems.

Keywords-component; Computer Science, Text Mining, mathematical foundations, applied computing, Robotics, Embedded Systems

12. Paper 31031543: d-HMAC — An Improved HMAC Algorithm (pp. 89-96)

Mohannad Najjar, University of Tabuk, Tabuk, Saudi Arabia

Abstract—The keyed-hash message authentication code (HMAC) algorithm is a security tool primarily used to ensure authentication and data integrity in information systems and computer networks. HMAC is a very simple algorithm, and relies on hash functions that use a secret key. HMAC's cryptographic strength is based on the use of effective cryptographic characteristics such as balancing and the avalanche effect. In this study, we develop a new algorithm, entitled dynamic HMAC (d-HMAC), to improve and enhance the cryptographic characteristics of HMAC. The improved algorithm provides stronger resistance against birthday attacks and brute force attacks. To achieve this objective, HMAC constant values *ipad* and *opad* are dynamically calculated in d-HMAC. Values for *ipad* and *opad* will be obtained from the HMAC input message, the public key of the receiver, and a substitution-box (*S*-box) table with enhanced security characteristics specifically created for this purpose. We demonstrate that the improved d-HMAC algorithm is more resistant to known cryptographic attacks, and prove that it exhibits similar or better cryptographic characteristics than HMAC.

Keywords-cryptography; data integrity; authentication; MAC; HMAC; hash functions; SHA-256

13. Paper 28021519: Hadoop Architecture and its Functionality (pp. 97-103)

Dr. B V Ramana Murthy,Jyothishmathi College of Engg and Technology, Shamirpet, India
Mr. V Padmakar, Guru Nanak Institutions Technical Campus, Hyderabad
Mr. M Abhishek Reddy, Jyothishmathi College of Engg and Technology, Shamirpet, India

Abstract - Hadoop is nothing but a "framework of tools" and it is a java based programming framework (In simple terms it is not software). The main target of hadoop is to process the large data sets into smaller distributed computing. It is part of the Apache project sponsored by the Apache Software Foundation. As we observe in database management system, all the data are stored in organized form by following the rules like normalization , generalizations etc., and hadoop do not bother about the DBMS features as it stores large amount of data in servers. We are studying about Hadoop architecture and how big data is stored in servers by using this tools and the functionalities of Map Reduce and HDFS (Hadoop File System).

Keywords: Big Data, HDFS, Map Reduce Task Tracker, Job Tracker, Data Node, and Name Node.

14. Paper 31031509: Data mining methodologies to study student's academic performance using the C4.5 algorithm (pp. 104-113)

Hythem Hashim (1) , Ahmed A. Talab (2) , Ali Satty (3) , and Samani A. Talab (1)
(1) Faculty of Computer Science and Technology, Alneelain University, Khartoum, Sudan.
(2) White Nile College for Science and Technology, White Nile state ,Kosti.
(3) School of Statistics and Actuarial Sciences, Alneelain University, Khartoum, Sudan.

Abstract - The study placed a particular emphasis on the so called data mining algorithms, but focuses the bulk of attention on the C4.5 algorithm. Each educational institution, in general, aims to present a high quality of education. This depends upon predicting the unmotivated students before they entering in to final examination. Data mining techniques give many tasks that could be used to investigate the students performance. The main objective of this paper is to built a classication model that can be used to improve the students academic records in Faculty of Mathematical Science and Statistics. This model has been done using the C4.5 algorithm as it is a well-known, commonly used data mining technique. The importance of this study is that predicting student performance is useful in many different settings. Data from the previous students academic records in the faculty have been used to illustrate the considered algorithm in order to build our classification model.

Keywords: Data mining, The C4.5 algorithm, Prediction, Classification algorithms.

15. Paper 31031533: Resource Modeling for the Development of a Decision-making System - Applied HCEFLCD of Morocco (pp. 114-118)

K. Oubedda, M. Khalfaoui, A. Ettahir
Systems Analysis Laboratory of Information Processing and Integrated Management (LASTIMI) School of Sale Technology, University Mohammed V Agdal

Abstract - The aim of this work and to develop a decision support system for the operation of a model including the main stakeholders of the High Commissioner for Water, Forests and Desertification Control (maker / managers, administrative, customers). This system is based on the relationship between the actors and their activities and their needs vary by contribution in time. It aims to make available to managers a set of dashboards that can improve the quality of provided services. We begin by modeling the actors up and clean process for studying both their organizations and their activities and needs. The first applications of this work has focused on data for the Directorate of Planning, Information System and Cooperation, and the Directorate of Forest Estate, Legal Affairs and Litigation. The results are encouraging.

Keywords: Information systems, decision support systems, dashboards, databases, modeling.

Where is the Cybersecurity Hero?

Practical Recommendations for Making Cybersecurity Heroism More Visible in Organizations

Paul D. Nugent, Ph.D. (corresponding author), Emilio Collar, Jr., Ph.D.

Management Information Systems, Ancell School of Business
Western Connecticut State University
Westside Classroom Building, Room 203
181 White Street, Danbury, CT 06810, USA

***Abstract*—Given the current state of cyber crime, cyber attacks, and cyber warfare, it is easy to argue that steps taken by those in cybersecurity and Information Assurance (IA) roles to thwart these attacks deserve heroic status. Yet, the reality is that these functions are rarely perceived by their organizations or by broader society as heroic and this has negative consequences for job satisfaction and for the attractiveness of careers in these fields. This paper explores the concept of the hero broadly as well as in organizational literature to understand how heroes are made in organizations and why the nature of cybersecurity and IA work present barriers to perceptions of heroism. This is because the intrinsic nature of security work focuses on vulnerabilities and therefore differs from other types of work that are focused on new capabilities. The paper concludes with practical recommendations on how managers, industry leaders, and educators can take steps to overcome these limitations and make careers in cybersecurity and IA more attractive.**

***Keywords*: Heroes; Heroism; Cybersecurity; Information Assurance; Information Security**

I. Introduction

This paper begins with an analysis of the hero in popular culture, sociology, philosophy, and organizational literature to think about heroism relative to cybersecurity and Information Assurance (IA) roles. We argue that there are unique factors that strongly discourage those occupying these organizational roles from being perceived by their coworkers or by society in general as being heroes even though their protection of individual and organizational well-being often crosses into heroic thresholds. The purpose of the paper is to understand these factors and to derive practical guidance for making this heroism more visible and for making careers in cybersecurity and IA more attractive.

II. The Hero

A. The Archetype of the Hero

To the psychologist Carl Jung, archetypes represent patterns of meaningful events (e.g., birth and death), figures (e.g., great mother, devil, wise old man, and hero), and motifs (e.g., creation and apocalypse) that are common across most cultures and historical epochs [1]. To Jung, archetypes are to intuitions (psychic apprehension) as instincts are to behaviors and are therefore deeply rooted in our psychology [2]. Although not amenable to scientific refutation, archetypes have been found to be useful in analyzing recurring patterns in mythology and in cultural studies [3] including heroes [4]. Furthermore, one need not go beyond one's own knowledge to consider how prevalent the hero figure is in ancient mythology, in history (e.g., wars, pioneering), and popular culture (e.g., literature and movies).

Philosophers and sociologists have also explored the role of the hero in society. To Emile Durkheim, the hero plays a central role in defining the sacred in ancient mythologies [5]. Friedrich Nietzsche's "superman" (overman) is heroic in the sense that he will create new values and new human potentialities in the absence of a divine authority [6]. In addition, Karl Marx considered the hero's self-sacrifice during times of war and revolution to be noble [7]. However the philosopher who wrote the most on heroism was G. W. F. Hegel who defined the hero as "somebody who artistically embodies the unity of individuality and universality in mythical or violent times" [8] and identified four representative heroes: Socrates, Alexander the Great, Caesar, and Napoleon.

And yet alongside these lofty mythical and historical "Heroes" with capital H's, there are also lesser, albeit still socially important, "heroes" with lowercase h's in our day-to-day lives. The key feature, whether it be self-sacrifice, protection, intelligence, creativity, or skill, is that the individual stands out from the crowd in a manner that goes beyond normal expectations and is highly appreciated by others. A parent can be a hero to a child, a political leader can be hero to her community, and a previously unsung employee can save his company from disaster with the right action at the right time. Therefore anyone, given the right circumstances, can become a hero and the archetype of the hero has implications for one's own participation and sense of meaningfulness in various social institutions.

In this light, studies show that there is variation in the ways in which individuals define heroes, but what is common is that their definitions of heroes draw upon specific features of the heroic figure as well as, more interestingly, modes of *identification* with the hero [9]. In other words, heroes and heroism have strong implications for how we view ourselves and our identity/status across various institutional spheres of life including our work life. The potential to be seen as a hero, then, if the circumstances arise, can then be framed as a deep source of motivation for individuals in work and non-work settings that should not be overlooked.

B. *Heroes in Organizations*

In modern times the institution within which man establishes his/her most significant social identity is the organization [10]. Furthermore, over the last three or four decades researchers have found it very fruitful to frame organizations not as rigid formal machines, but rather as complex cultures with myths, ceremonies, rites, symbols, and heroes [11]. In their influential 1982 book *Corporate Cultures: The Rites and Rituals of Corporate Life*, Deal and Kennedy claim that the hero is an individual who is particularly important to the culture of the organization because they assert, "If values are the soul of the [organizational] culture, then heroes personify those values and epitomize the strength of the organization [12, p. 37]. Furthermore, they claim heroes produce many positive outcomes or impacts on organizations:

- make success attainable and human
- provide role models
- symbolize the company to the outside world
- preserve those values that make companies special or meaningful
- set standards of performance
- motivate employees
- encourage greater individual persistence in striving for organizational goals
- They urge people to identify their own personal success with the organization's success [12]

Leadership studies and managers came to adopt the concept of the hero as a variable to be manipulated within organizations to improve performance. In the same vein as the seminal work "The One Minute Manager," managers now had an explicit goal to identify heroes, communicate their heroism to the rest of the organization, and publically reward them [12].

However, as with many organizational interventions/programs, these efforts often have unintended or negative consequences because they frame organizational culture from a top-down managerial point-of-view that ignores other realities such as subcultural differences, social complexity, and ambiguity [13]. For example, in the study *"Please Don't Make Me a Hero": A Re-Examination of Corporate Heroes*, the authors draw on empirical data to show that these formal interventions often lead to negative stigma and embarrassment on the part of those identified as heroes and also to the resentment of coworkers that felt they should have shared in the rewards [14]. They recommend that instead of explicitly trying to manipulate hero status within organizations, "executives can accomplish more by working on perspective, praise, and trying to influence the daily practice in their organizations than by trying to create heroes" [14].

Nonetheless, while attempts by management to deliberately "manage" heroes within organizations, as Deal and Kennedy prescribe, may be problematic because they ignore some social realities, to the extent that heroes do arise of their own accord within informal organizational cultures, the positive outcomes previously mentioned seem reasonable and desirable. The goal, then, is to facilitate the creation of heroes within an organization, but to do so in a manner that more fully acknowledges and accounts for organizational culture at deeper levels.

The pivotal question in this paper is, then, what happens when employees (such as those performing cybersecurity or IA duties) in the organization act in a manner that fits the definition of hero but for various reasons, that we will explore, are not identified as such by the formal or informal organization? What steps can managers take to create these heroes while avoiding the aforementioned pitfalls? What implications does this have for the organization and, perhaps more importantly, for job satisfaction or career attractiveness?

III. The Unique Nature of Cybersecurity and Information Assurance Work

The etymology of the word hero reveals that the word derives from meanings of *defender* and *protector* [15]. This is not surprising, as historically heroes were likely those who protected the wellbeing of communities in times of crisis such as famine, draught, and war. In the current information age the wellbeing of individuals (finances, personal information, etc.) as well as organizations (assets, secrets, availability of services, etc.) require protection and those in the roles of cybersecurity and IA perform the difficult role of the defender and the protector. Why, then, are these defenders and protectors rarely hailed as heroes by individuals or by organizations?

A. *Visibility of Threats, Mitigations, and Outcomes*

One explanation is that their successful protections are not recognized or recognizable. When a village or community is under attack in a "conventional" war, observers have a clear understanding of the threat (the enemy), the means of defense, and whether or not the attack was successful. It is a straightforward matter, then, to identify responsible groups or individuals if the attack is unsuccessful. In the world of networked (i.e., Internet) communications, however, such determinants are much less visible or obvious to the general observer. There is not a single potential "attacker" and there is a broad spectrum of motivations behind the attacks. Therefore the threat to the organization is not monolithic or clearly identifiable to the general organization. In most cases even the cybersecurity employees are not aware of specific attackers but must be prepared to defend against a wide array of potential attack categories. The enemy, then, has no *face* and this renders it more difficult for members of the organization to perceive successful defenses as heroic.

In addition to the anonymous nature of the attacks, the manner in which the protectors defend against the attacks through controls and mitigations also contributes to the difficulty in viewing them as heroes. Most organizational members and the general public who depend upon the defense of these assets to protect their wellbeing have very little knowledge of the kinds of mitigations and controls that are skillfully put into place. They likely do not understand the nature of gateways, encryption, host-based security systems, patches, etc. and therefore do not appreciate the protective work as a whole.

Also, most organizations depend upon the Internet for marketing and customer service and this means connectivity to the Internet and a demilitarized zone buffering internal networks from external Email and Internet messaging. However as long as individuals in the internal network can access the Web and receive external Emails to perform their duties, there are non-mitigatable social engineering threats. In other words, in a risk management framework the organization *accepts* a balance of risk and protection and that the potential for the exploitation of some vulnerabilities will persist. It is difficult to be viewed as a hero when you consciously accept such vulnerabilities.

Another factor that makes cybersecurity protection different than the war analogue is the level of certainty of whether an attack occurred and/or was successfully thwarted. While statistics on certain kinds of attacks can be gathered and analyzed, many other kinds of attacks are difficult to identify with any level of certainty [16]. Certainty, rather, is reserved for successful attacks and in a context of failures to protect the wellbeing of the organization or its customers; heroes rarely emerge.

Furthermore, even when it is clearly understood that a potentially grave attack occurred and it was successfully defended against, there is the problem of attribution of credit. Was an individual or group truly heroic in identifying and thwarting the attack, or were they simply doing their routine job and implementing industry best practices for security mitigation/control? IA and cybersecurity practices are highly complex and encompass many different topical domains (e.g., CISSP domains) and to organizational members not familiar with these practices, they are seen as a confusing “black art.” Therefore, while those close to “the action” will be able to attribute successful protection to something akin to heroism or merely “doing one’s job,” others in the organization, including most managers, are not able to make these attributions.

B. Creator versus Protector

Science, engineering, and software development enjoy the positive image of discovering, creating or inventing things that people find important or useful. Within the broader culture or within individual organizations, it is common to find heroes associated with these roles [17]. These heroes are not protectors or defenders as we discussed in the previous section, but rather are heroic because they employ their knowledge, creativity, vision, and hard work to *produce* something new. From early on in their education, students of these disciplines come to identify strongly with this facet of their work and the potential to be admired and respected for a new technology or a new capability they produce.

Now consider the nature of cybersecurity and IA. Rather than putting their efforts toward creating capability, the cybersecurity employee is focused on protecting against vulnerabilities inherently associated with capabilities. In larger organizations there is a division of labor between those who *design* the system and those who *protect* the system and those in technical cultures tend to reserve prestige and praise for the more visible and perhaps more understandable design side.

C. Attacker versus Defender

Since before the rise of the Internet there has been a cultural fascination with computer and network hacking/cracking. Whether it is decrypting communication cyphers during wartime, or penetrating a “digital fortress” [18], or breaking into government networks, or con/fraud artists such as Frank Abagnale dramatized in the movie *Catch Me If You Can*, the attacker has attained a hero status bordering on romantic [19].

It is also evident from online hacking communities/forums that hackers freely brag about their successful attacks with impunity and perceive their skills as virtuous [20]. Furthermore, their skills were also in great demand from communities hoping to protect their networks and notorious hackers and hacking groups were being hired to protect organizations. For example in the article *When Hackers Become Heroes*, the author claims that a shift occurred in May of 1998 when a groups of hackers named “L0pht” based in Boston testified before the U.S. Senate and they began to be perceived less as criminals and more as possessing skills that were extremely valuable in understanding how to protect systems [21].

The key point is that it is, ironically, the attacker and not the defender that achieves hero status in these contexts. Even though these skills may be translated into protective roles, it is nonetheless the attacker/hacker/cracker that is initially perceived as the hero. In contrast, the day-to-day role of the cybersecurity or IA employee, even when successfully staving off grave attacks, is not often portrayed as heroic in popular culture.

IV. Attractiveness of Cybersecurity Careers

The preceding analysis contributes toward our understanding of the reasons why cybersecurity jobs and careers may be perceived as less attractive to potential aspirants. In addition to these are factors such as pay and the technical climate of the work itself.

For example, one trade article considers how despite it being in high demand, college students trained in computer science are discouraged from careers in cybersecurity because it is not seen as "hip" and initial pay levels are not at the levels of alternatives such as computer engineering or software development [22].

Another study highlights how, even from the point of view of highly skilled hackers that have been hired into cybersecurity roles, the work is described as, "uncreative, bureaucratic, and restrictive [23]." Other authors have pointed out that younger candidates are simply unaware of cybersecurity and IA careers and, "In summation, the problem is that millennials either haven't heard of careers in cybersecurity, or, if they have, it sounds like a boring and potentially unethical boy's club. Not a great combination [24]."

Finally, for those who do explore these careers in depth, it becomes clear that it is difficult to secure the higher paying jobs without industry certifications and without numerous years of experience [25].

V. Making Cybersecurity and IA Careers more Attractive

There are many steps managers, industry leaders, and educators can take to overcome the factors that render careers in cybersecurity less attractive. For example entry-level salaries could be increased, the work could be made less bureaucratic and restritive, and colleges could do more to "sell" these as potential careers. Yet even with these efforts, the assumption of this paper is that social factors discouraging those in these organizational roles from being identified and hailed as heroes where this status is objectively justified need to be addressed.

In the trade literature there are some, although not many, attempts to do this. For example one article uses the James Bond analogy to associate the work with an exciting international spy context [26]. While this may be overly romanticizing the true nature of the work, there is still a grain of truth in it when one considers the ever-present potential of large organizations being targeted by activist groups or political enemies throughout the world. Others have attempted to highlight the nobility of being a protector in hopes to make these careers more attractive to women and minorities [27] or highlight the work of cybersecurity in insurance contexts [28].

Missing from these gestures are more direct ways to make these careers more attractive such as highlighting the heroic or potentially heroic nature of the work itself in a social context. It is here that we need to return to the earlier analysis of heroism in organizational contexts for guidance. The starting point, then, must be the assumption that beyond "hygiene" factors such as salary, workers desire meaningful work and meaningfulness does not derive merely from the individual nature of these roles, but moreso the social and cultural framings of these roles. Within organizations and within organizational fields professional desire status in a group where their work, knowledge, and skills are appreciated and valued.

For example, earlier in the paper we argued that some of the factors discouraging this status stem from limitations on the visibility of threats, mitigations, and successful outcomes. Rather than attempting to single out individuals as heroes as is Deal and Kennedy's prescription, we recommend instead to take steps toward making these phenemena more visible. For instance, managers can work with with the cybersecurity/IA group within the organization to translate what is known in department spreadsheets and databases about threats, mitigations, and outcomes to be interpretable by a broader organizational audience. Much of this "objective" information will be readily available in larger organizations based on risk management and risk assessment tasks but are not typically translated to a general audience. Therefore with minimal effort, this information can be conveyed through security bulletins and organizational intranets so that the organizational as a whole can come to be aware of the objective nature of threats, the efforts being undertaken to protect against the threats, and success stories.

In this manner the potential negatives discussed earlier in *Please Don't Make me a Hero* are overcome because no one individual is being singled out. Rather, the members of the groups enjoy elevated status within their organization because their protective roles are being appreciated and may even reach heroic levels where threats are palpable and concrete steps taken by the group are seen as successful. Unfortunately there are practical forces working against this approach as publicizing details about specific attacks could potentially reveal the organization's protection strategies making them more vulnerable to attackers. The challenge to managers is to highlight the graveness of the threats and the effectiveness of their protections without revealing too many of the technical details.

Furthermore, the bias toward capability (creativity/production) and away from vulnerability (security) can be addressed. For example, one journal article attempts to show that, from the point of view of the philosophy of technology, *vulnerabilility* is just as inherent to complex systems as their *capabilility* implying that from a work studies perspective addressing vulnerability should be given equal social status [29]. Therefore managers can do more to acknowledge the nature of the systems that they create and that with additional capability comes an equivalent, or perhaps greater, level of vulnerability that needs to be properly understood and managed. This may require incentivizing engineers and computer scientists that have been successful in designing systems to move to the "other side" in roles to protect those systems. The goal would ultimately be to elevate those departments focused on security to the same organizational status as those tasked with design responsibilities. In concert with this, management can leverage the hacker-as-hero identity and have the employees conduct more "ethical hacking" through various forms of penetration testing.

Finally, although not addressed directly in the preceding analysis, there are steps that educators can take to make careers in cybersecurity and IA more attractive to their students. On a practical level, with ever increasing network complexity and threat environments, cybersecurity roles are expected to be in high demand for the foreseeable future as the demand for other potential design-related careers may be flat-lining or declining. More in keeping with the overall thrust of this paper, however, is the need for educators to impart to students in the classroom the objective nature of cybersecurity and IA work and the central role that this work plays in protecting the welfare of individuals, organizations, and governments.

REFERENCES

[1] C. G. Jung, quoted in J. Jacobi, *Complex, Archetype, Symbol*, p. 114 London, 1959.

[2] C. G. Jung, *Psychological Types*, Bollingen Series, Princeton, 1971.

[3] R. Johnson, *We: Understanding the Psychology of Romantic Love*, Harper, 1983.

[4] J. Campbell, *The Hero with a Thousand Faces*, Princeton: Princeton University Press, 1968.

[5] M. Mauss, H. Hubert, A. Riley, and S. Daynes, *Saints, heroes, myths, and rites: classical Durkheimian studies of religion and society*, Paradigm Publishers, Boulder, CO, 2009.

[6] R. Luyster, "Nietzsche/Dionysus: Ecstasy, Heroism, and the Monstrous," in *Journal of Nietzsche Studies*, No. 21, Spring, pp. 1-26. Penn State University Press, 2001.

[7] A. Wood*, Karl Marx – Arguments of the Philosopher*, Psychology Press, 2004.

[8] J. Früchtl, *The Impertinent Self: A Heroic History of Modernity*, p. 36, Stanford University Press, 2009.

[9] M. Sullivan and A. Venter, "Defining Heroes Through Deductive and Inductive Investigations," in *The Journal of Social Psychology*, 2010.

[10] M. Weber, "Bureaucracy," in *Classics of Organization Theory*, Shafritz, J. M. & Ott, J. S. (eds.) 3rd Ed. Brooks/Cole Publishing Co., 1973.

[11] J. W. Meyer and B. Rowan, "Institutionalized Organizations: Formal Structure as Myth and Ceremony," in *American Journal of Sociology*, 83, No. 2 (September), 340-63, 1977.

[12] D. Kennedy, *Corporate Cultures: The Rites and Rituals of Corporate Life*, 1982.

[13] J. Martin, *Cultures in Organizations: Three Perspectives*, Oxford, 1992.

[14] A. L. Wilkins, "Please Don't Make Me a Hero: A Re-Examination of Corporate Heroes," in *Human Resource Management*, Volume 29, Issue 3, pages 327–341, Autumn (Fall), 1990.

[15] Online Etymology Dictionary. http://www.etymonline.com

[16] *Kaspersky Security Bulletin*, http://media.kaspersky.com/pdf/KSB_2013_EN.pdf, (2013). Accessed Retrieved 28 June 2014

[17] Z. Pascal, "Where Are Today's Engineering Heroes? By failing to celebrate its finest contributors, the profession risks far more than mere obscurity," in *IEEE Spectrum*, June, 2014.

[18] D. Brown, *Digital Fortress*, St. Martin's Press, 1998.

[19] S. Levy, *Hackers: Heroes of the Computer Revolution*, 1984.

[20] Q. Campbell and D. Kennedy, "The psychology of computer criminals," In Bosworth, et al (Eds.), *Computer security handbook*, New York, NY: John Wiley & Sons, Inc., 2009.

[21] M. Rogers, "When hackers became heroes," in *Special for USA Today*, April, 2014.

[22] *Mashable.com*. For Job Security, Try Cybersecurity, Experts Say. http://mashable.com/2012/05/29/cybersecurity-career/. 2012. Accessed 1 July 2014

[23] W. Jackson, "The pros and cons of government cybersecurity work," in *Cybereye*. http://gcn.com/Articles/2010/08/23/Cybereye-cybersecurity-jobs.aspx?Page=1. August, (2010). Accessed 1 July 2014.

[24] B. Richmond, "The Cybersecurity Industry Is Hiring, But Young People Aren't Interested," http://motherboard.vice.com/blog/the-cybersecurity-industry-is-hiring-but-young-people-arent-interested. October, (2013). Accessed 6 June 2014.

[25] *SANS Institute*, "Cybersecurity Professional Trends: A SANS Survey," in *InfoSec Reading Room*, This paper is from the SANS Institute Reading Room site, May, 2014.

[26] *Jobs.net*, Cyber Security – The Best & Worst of a Modern James Bond Gig. http://www.jobs.net/Article/CB-19-Talent-Network-IT-Cyber-Security-150-The-Best-amp-Worst-of-a-Modern-James-Bond-Gig. (2014). Accessed 2 July, 2014.

[27] C. Madren, "Cyber Heroes in Training," in *Washington Post,* August, 2013.

[28] S. Groves, "The Unlikely Heroes of Cyber Security: Viruses, Privacy Breaches, and Other Malicious Cyber Activity Regularly Threaten Organizations' Vital Information," in *Information Management*, Vol. 37, No. 3, 2003.

[29] P. D. Nugent and A. Ali, "Frankenstein's other Monster: Toward a Philosophy of Information Security," in *International Journal of Computer Science and Information Security*, Vol. 10, No. 4, 2012.

User Customizable Privacy-Preserving Personalized Web Search

Akhila G.S
M.Tech student
Department of Computer Science and Engineering
Mohandas College of Engineering
Anad, Trivandrum, Kerala, India

Mr.Prasanth R.S
Asst.Professor
Department of Computer Science and Engineering
Mohandas College of Engineering
Anad, Trivandrum, Kerala, India

***Abstract*— Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. This paper proposes a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Present an algorithm, namely GreedyIL, for runtime generalization. The experimental results show that GreedyIL performs efficiently.**

Keywords—Personalized Web Search,User profile

1 INTRODUCTION

Nowadays, computers and internet has become inseparable parts of our life. Throughout the world, web has become the best source of useful information. Search engines play a key role in finding out the information. They are enhanced with new advanced search technologies. Though search engines find much information with one key word, fail to provide the accurate and exact data that is required. Hence, the most significant point in the applications of the search engines is to find accurate information immediately. This aspect of accurate and immediate information for a search can be solved by personalized web environments. Personalized web search is a general category of search techniques aiming at providing better search results. The solutions to PWS can generally be categorized into two types, namely *click-log-based* [1] and *profile-based* [1] methods. In click-log based methods, they simply impose preference to clicked pages in the user's query history. One limitation to reduce its applicability is that it can only work on repeated queries from the same user. The profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques.

A user profile contains the personal information or interests of a particular person. Different profiling techniques are available to construct the user profile. Before the user profile construction a system needs to identify the interests of users. The sources we have used in constructing a user's profile are: bookmarks from a social bookmarking site, web communities, social networking services, blogs of interests etc. The first step in the construction of user profile is that pre processing. The pre processing step involves stop word removal and stemming. These are then converted to feature vectors where the features are the terms in the documents after the pre processing step. After performing any clustering algorithm, we get several clusters and clusters would represent interests. So if we assign weight ages to interest vectors on the basis of documents downloaded and browsed we get a fairer representation of a user's current interest. As far as weight ages are concerned they can be assigned proportional to the number of documents assigned to each cluster on the basis of the similarity metric. So user profile has very important role in effectiveness of search quality.

2 RELATED WORKS

A. Pretschner and S. Gauch proposed personalized web search based on ontology. In this technique the authors created user profile by analyzing surfed pages to identify their content and the time that was spent on it. When pages about certain subjects are visited again and again, this is taken as an indication of the user's interest in that subject. The main advantage of this approach was that except for the act of surfing, no user interaction with this system is necessary. But disclose the user private information was the main problem in this approach. Because here the user profile created based on user's surfed pages and created user profile could viewed by publically accessible search engine.

L. Fitzpatrick and M. Dent developed personalized web search, in this users have to register personal information such as their interests, age, and so on during the training period. Another one method was users have to provide feedback on relevant or irrelevant judgements, by rating on a scale from 1 to 5. Here 1 indicates very bad and 5 indicates very good. This approach had a lot of limitations. Explicit construction of user profiles has several drawbacks. Sometimes user provides inconsistent or incorrect information, it affect the construction of user profile. It was a time consuming process.

K. Sugiyama, K. Hatano, and M. Yoshikawa suggested an adaptive web search based on user profile. The main advantage of this approach was user profile constructed without any effort or feedback from user. The main problem in previous approach was need continuous user interaction. This technique solved that problem. In this system, when a user submits a query to a search engine via web browser, the search engine returns search results corresponding to the query.

Based on the search results, the user may select a web page in an attempt to satisfy his/her information need. In addition, the user may access more web pages by following the hyperlinks on his/her selected web pages and continue to browse. The proposed system monitors the user's browsing history and updates his/her profile whenever his/her browsing page changes. When the user submits a query the next time, the search results adapt based on his/her user profile. Fig.1. shows the system architecture. Here also the main problem was discloses of user privacy. It gave better performance than previous techniques, but the created user profile completely accessible by search engine.

M. Spertta and S. Gach recommended personalized web search technique based on user's search history. In this approach, the authors constructed user profile by analysing user's search history. Search engines index millions of documents on the Internet and allow users to enter keywords to retrieve documents that contain these keywords. Browsing is usually done by clicking through a hierarchy of subjects until the area of interest has been reached. In this approach is based on building user profiles based on the user's interactions with a particular search engine. For this purpose, the developers implemented GoogleWrapper. A wrapper around the Google search engine, that logs the queries, search results, and clicks on a per user basis. This information was then used to create user profiles. A wrapper for Google that implicitly collects information's from users. Users register with their email addresses in order to create a cookie storing their user id on their local machines. When user submits a query, GoogleWrapper logs the query and the userID and then forwards the query to the Google search engine. The search engine returns result back to the user based on the created user profile of particular userID. To keep users secrets was the major problem in this approach.

Xuwei Pan, W. Zhengcheng and G. Xinjian proposed context-based adaptive personalized web search for improving information retrieval effectiveness. In this approach, the authors proposed a novel adaptive personalized technique based on context to adapting search outputs consistent with each user's requirement in different situations for relevant information with slight user effort. Experimental observations prove that the adaptive personalized search system is executed by most of users and the approach to personalize web search is effective.

Y. Chen, H. L. Hou and Z. Yan-Qing recommended a personalized context-dependent web search agent using semantic trees. In web searching applications, contexts and users' preferences are two significant features for Internet searches in some way that outputs would be much more appropriate to users' requests than with existing search engines. Researchers had planned a concept-based search agent which utilizes conceptual fuzzy set (CFS) for matching contexts-dependent keywords and concepts. In the CFS model, a word accurate meaning may be determined by other words in contexts. Owing to the fact that various combinations of words may become visible in queries and documents, it may be complicated to identify the relations between concepts in all possible combinations. To avoid this problem, the authors proposed a semantic tree (ST) model to identify the relations between concepts. Concepts are symbolized as nodes in the ST, and relations connecting these concepts are represented by the distances between nodes. Furthermore, this paper makes use of the users' preferences for personalizing search results. Finally, the fuzzy logic will be utilized for finding which factor, semantic relations or users' preferences will control results.

L. Fang, C. Yu and W.Meng recommended personalized web search for improving retrieval effectiveness. In this paper, the authors propose a novel approach to learn user profiles from users' search histories. The user profiles are then utilized to enhance retrieval efficiency in web search. A user profile and a common profile are studied from the search history of the user's and a category hierarchy, respectively. These two profiles are integrated to map a user query into a group of categories which corresponds to the user's search intention and provide a context to disambiguate the words in the user's query. Web search is performed according to both the user query and the group of categories. A number of profile learning and category mapping approaches and a fusion algorithm are presented and evaluated.

D. Chen, J.C. Patra and F.C. Peng suggested personalized web search with self-organizing map. The commonly used web search engines provide the similar answer set for different preferences. Personalized web search performs the search for all users according to their preference. With the intention of minimizing the consumption of time on browsing irrelevant documents, this paper suggests an intelligent Personal Agent for Web Search (PAWS). The PAWS cleverly utilizes the self organizing map (SOM) as the user's profile and therefore, is capable of providing high quality answer set to the user.

Y. Xu, K. Wang, B. Zhang, and Z. Chen proposed a privacy-enhancing personalized web search. Users are uncomfortable with exposing private preference information to search engines. Fig.2. provides a system overview of proposed system. In this approach, authors introduce an algorithm which automatically builds a hierarchical user profile that represents the user's implicit personal interests. Only portions of the user profile will be exposed to the search engine in accordance with a user's own privacy settings. A search engine wrapper is developed on the server side to incorporate a partial user profile with the results returned from a search engine. Rankings from both partial user profiles and search engine results are combined. The customized results are delivered to the user by the wrapper. In this approach, user profile kept as exposed plus private. That is why user's privacy could be preserves somewhat. This profile-based PWS does not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user without making distinctions. Here ranking is take place based on only exposed user profile.

B. Smyth proposed a community-based approach to personalizing web search. Researchers can influence the underlying knowledge produced within search communities by gathering user's search behaviours- the queries they enter and results they choose- at the community level. They can

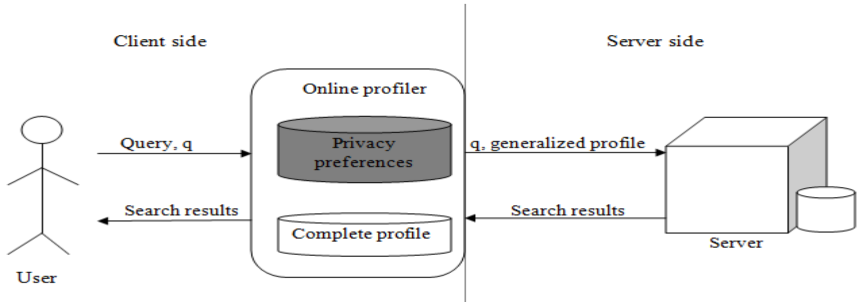


Figure 1: Architecture of UPS

make use of this data to construct a relevance model that provides the promotion of community-relevant results throughout standard web search. This paper focuses on the collaborative web search technique that encourages the suggestion that community search behaviours can offer valuable form of search knowledge and sharing of this knowledge makes adapting conventional search-engine outputs possible.

3 Problems in previous works

Many personalized web search approaches have been discussed in the literature survey. The problems with the existing methods are explained in the following observations:

1. *The existing profile-based PWS do not support runtime profiling.* A user profile is typically generalized for only once offline, and used to personalize all queries from a same user without making distinctions. Such “one profile fits all” strategy certainly has drawbacks given the variety of queries.

2. *The existing methods do not take into account the customization of privacy requirements*. This means that some user privacy to be overprotected while others insufficiently protected. Unfortunately, few prior works can effectively address individual privacy needs during the generalization.

3. *Privacy* is another one major problem in the existing approaches.

4. *Unnecessary exposure of the user profile*. Sometimes users really not need to personalize a particular query. The existing system personalizes all queries submitted by user based on user profile or click-through data.

5. *Many personalization techniques require iterative user interactions when creating personalized search results*. They usually refine the search results with some metrics which require multiple user interactions.

4 System architecture

From the literature survey it is concluded that most of the PWS system primarily focused on to improve search quality. The profile-based PWS do not support runtime profiling. Also these methods do not take into account the customization of privacy requirements. Unnecessary exposure of the user profile is another one problem.

Propose a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements. The key component for privacy protection is an online profiler implemented as a proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified privacy requirements represented as a set of sensitive-nodes.

The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows:

- When a user issues a query on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile satisfying the privacy requirements.
- The query and the generalized user profile are sent together to the PWS server for personalized search.
- The search results are personalized with the profile and delivered back to the query proxy.
- Finally, the proxy either presents the raw results to the user, or re ranks them with the complete user profile.

5 Preliminaries

In this section, we first introduce the structure of user profile in UPS. Then, we define the customized privacy requirements on a user profile. Finally, we formulate the problem of privacy preserving profile generalization.

5.1 User profile

Here, each user profile in UPS adopts a hierarchical structure. Moreover, our profile is constructed based on the availability of a public accessible taxonomy, denoted as R. This repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic *t*, a corresponding node can be found in R, with the sub tree *subtr (t, R)* as the taxonomy accompanying *t*. The repository is regarded as publicly available and can be used by anyone as the background knowledge. Examples for such repositories are ODP, Wikipedia, WordNet and so on.

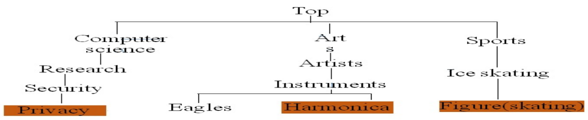


Figure 2: Sample user profile

A user profile H, as a hierarchical representation of user interests, is a rooted sub tree of R. Given two trees *S* and T, *S* is a rooted sub tree of T if *S* can be generated from T by removing a node set X from T, i.e., S= rsbtr(X,T). A diagram of a sample user profile is illustrated in Figure 1.

5.2 Customized privacy requirements

Customized privacy requirements can be specified with a number of topics in the user profile, whose disclosure to the server introduces privacy risk to the user. Given a user profile, the sensitive nodes are a set of user specified sensitive topics, whose sub trees are nonoverlapping.

The most straightforward privacy preserving method is to remove sub trees rooted at all sensitive-nodes whose sensitivity values are greater than a threshold. Such method is referred to as *forbidding*. We exemplified the limitation of the forbidding operation in next section.

5.3 Generalizing user profile

Now, we exemplify the inadequacy of forbidding operation. In the sample profile in Figure 2, *Figure* is specified as a sensitive node. Thus, rsbtr(S, H) only releases its parent *Ice skating*. Unfortunately, an adversary can recover the sub tree of *Ice skating* relying on the repository, where *Figure* is a main branch of *Ice skating* besides *Speed*. If the probability of touching both branches is equal, the adversary can have 50 percent confidence on *Figure*. This may lead to high privacy risk if *sen*(*Figure*) is high. A safer solution would remove node *Ice skating* in such case for privacy protection. In contrast, it might be unnecessary to remove sensitive nodes with low sensitivity. Therefore, simply forbidding the sensitive topics does not protect the user's privacy needs precisely.

To address the problem with forbidding, we propose a technique, which detects and removes a set of nodes X from H, such that the privacy risk introduced by exposing G= rsbtr(X,H) is always under control. Set X is typically different from S. This process is called *generalization*, and the output G is a *generalized profile*. For this purpose introduced GreedyIL generalization algorithm. Details of this algorithm will be presented in section 6.

6 UPS procedure

Specifically, each user has to undertake the following procedures in our solution,

1. Offline profile construction
2. Offline privacy requirements customization
3. Online query-topic mapping
4. Online generalization.

Offline-1. Profile construction. The first step of the offline processing is to build the original user profile in a topic hierarchy that reveals user interest. For constructing the user profile we need to track the user's search history. For each click to the links, the corresponding url information stored. Using these information and WordNet information, we can construct user profile. In section 7, we present details about this step.

Offline-2. Privacy requirement customization. This procedure first request the user to specify a sensitive-node set.

Online-1. Query-topic mapping. Given a query q, the purpose of query-topic mapping is to compute a rooted sub tree, which is called a seed profile, so that all topics relevant to q are combined in it. For example, by applying the mapping procedure on query "Eagles", we obtain a relevant set T (Eagles) as shown in Table 1. Overlapping the sample profile in Figure 1 with its query-relevant trie R (Eagles) gives the seed profile G_b, whose size is significantly reduced compared to the original profile.

Topics in T(Eagles)
Top/Arts/Music/Artists/Eagles
Top/Sports/American football/NFL/Philadelphia Eagles
Top/Science/Biology/Animals/Birds/Raptors/Eagles
Top/Society/Military/Aviation/Aircraft/fighters/F-15/Eagles

Table 1: Contents of T(Eagles)

Online-2. Profile generalization. This procedure checks whether this seed profile satisfies privacy requirements of user when it is exposed to the server. In addition, this procedure computes the discriminating power for online decision on whether personalization should be employed.

7 GREEDYIL ALGORITHM

The greedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings. One important finding is that any prune-leaf operation reduces the discriminating power of the profile. Considering the process of pruning leaf t from G_i to obtain G_{i+1} in the ith iteration, maximizing DP (q, G_{i+1}) is equivalent to minimizing the incurred information loss.

The following description gives a brief idea about the GreedyIL algorithm. This algorithm takes seed profile, query, and privacy threshold as inputs. The output we get from this algorithm will be a generalized profile which satisfying user specified privacy threshold value. We maintain a priority queue of candidate prune-leaf operators in descending order of the information loss caused by the operator. First, we check whether the submitted query is distinct or not. If it is a distinct query then there is no need to perform pruning operations. If not, we have to do some operations.

1. Obtain the seed profile G_i from online-1.
2. Insert each topic and corresponding information loss into Q for all topics in seed profile.
3. Check whether this seed profile exposed to server will make any privacy risk. If so perform following operations, otherwise go to step 9.
4. Pop a prune-leaf operation on t from Q.
5. Set s as it parent. Perform pruning operation.
6. Check if t has no sibling, and then just insert s and corresponding information loss to Q. Then go to step 8.
7. If t has siblings,
    7.1. Merge t into shadow-sibling.
    7.2. Check if these t's siblings have any operations in Q. If not, just insert s and corresponding information loss into Q. Otherwise, updates the IL- values for all operations on t's siblings in Q.
8. Update i.
9. Return this G_i as generalized profile G^*.

After performing these operations, we will get a generalized user profile in runtime with the light of query terms. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search. The search results are personalized with the profile and delivered back to the query proxy. The proxy presents the raw results to the user.

8 IMPLEMENTATION

The UPS framework is implemented on a PC with an Intel Core i5 2.67-GHz CPU and 4-GB main memory, running Microsoft Windows 7. All the algorithms are implemented in Java. The topic repository uses the WordNet. First step in our thesis work was download dataset from web.

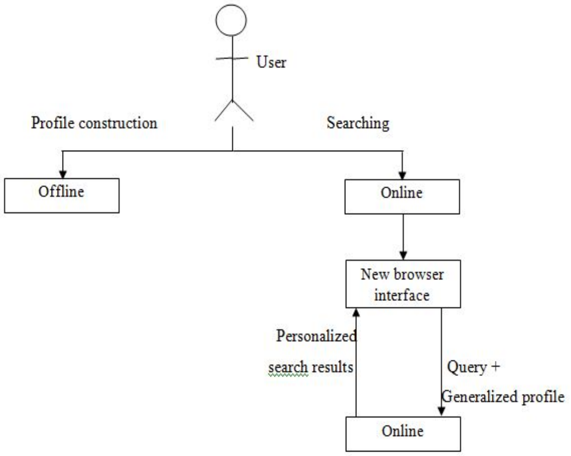


Figure 3: Implementation flow

We had a database called searchengine, which contains four tables. The first table "*urlinfo*" contains the urls, which was already visited by the users. This table also contains the most frequent words in each url. The second table named "*allwords*", gives the most frequent words in url already visited by users. The third table "*related_words*", which stores each frequent word search by the users, and corresponding related words in WordNet. The last table "*hide_words*", stores the words that are hidden from the public server.

We can divide the overall procedure of the thesis into three modules.

1. *Profile construction*: This is an offline process for identifying user's interests, for constructing user profile that we need to track the user's search history. So, the first form gives an interface for the user to search their queries. For each click to the links, the corresponding url information stored into the table "*urlinfo*". Internally, we calculate the most frequent words in each url. After performing stemming, count each word, if the count is more than predefined value it will be stored into the corresponding entry in the table. These words are also stored into the table "*allwords*". Now we have frequent words, which might be the interested topics of the user, also we need to find corresponding related words in the WordNet. For that we use "*allwords*" table's information and dataset. Trace all related words and store those words into "*related_words*" table. Based on all these tables we constructed the user profile in a hierarchical structure. We used simple tree construction java code for the hierarchical structure.
2. *Privacy requirement customization*: This procedure first requests the user to specify a sensitive-node set by selecting topics from hierarchical user profile. Those words are stored in the "*hide_words*" table.
3. *Query processing*: This procedure is an online processing. After creating user profile, users can directly search through new browser interface. When user submits a query, internally perform our GreedyIL algorithm to generalize the user profile with runtime in the light of submitted query. The query and the generalized profile are sent together to the PWS server for personalized search. The

search results are personalized with the profile and delivered back to the user.

9 Problems in base paper

The main disadvantage of base paper is there may be a chance of eavesdropping when generalized profile forwarded to the server. Based on generalized profile, the attacker will attempt to touch the sensitive nodes of the user by recovering the segments hidden from the original user profile, and computing a confidence for each recovered topic, relying on the background knowledge in the publicly available taxonomy repository. Figure 4 shows the attack model of PWS.

10 Conclusion and future work

Personalized Web Search is a promising way to improve search quality. Using this new search engine we got the personalized search results without revealing any personal information. We compared the search results from both Google and our new search engine. From that we understood that search results from search engine were more personalized and moreover we can save our valuable time. Search results from our search engine is same as that of search results from Google account or MyYahoo!, but more personalized. We need to register our personal information or interests before search through Google account or MyYahoo!, for getting personalized search results. Overall using this new search engine we provide privacy to the PWS.

For future work, we will try to resist adversaries with cryptographic concepts. We couldn't use the usual encryption techniques, because need to decrypt at the server side. The solution to this problem is applying homomorphic encryption, which allows computations to be carried out on ciphertext also generate an encrypted result which, when decrypted, matches the results of operations performed on the plaintext.

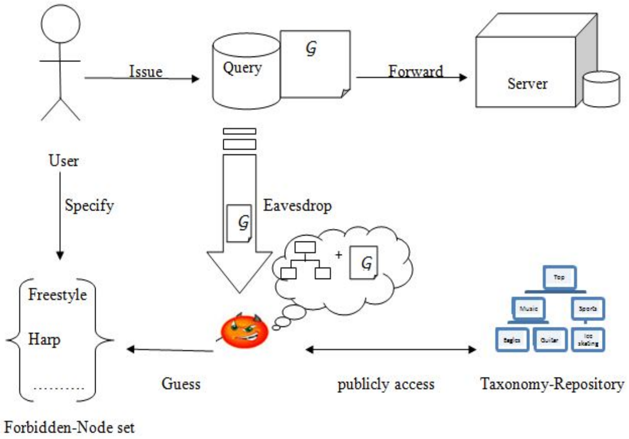


Figure 4: Attack model of PWS

References

[1] L.Shou,H. Bai,K.Chen, and G. Chen,"Supporting Privacy Protection in Personalized Web Search",vol.26,February 2014.

[2] A.Pretschner and S.Gauch,"Ontology-Based Personalizaed search and Browsing," Proc. IEEE 11th Int'l Conf.Tools with Artificial Intelligence,1999.

[3] L. Fitzpatrick and M.Dent,"Automatic Feedback Using Past Queries:Social Searching?," Proc.20th Conf.,1997.

[4] K.Sugiyama, K. Hatano, and M. Yoshikawa,"Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc.13th Int'l Conf. World Wide Web (WWW), 2004

[5] M.Spertta and S.Gach,"Personalizing Search Based on User Search Histories,"Proc.IEEE/WIC/ACM Int'l Conf.Web Intelligence,2005.

[6] F.Liu,C.Yu and W.Meng,"Personalized Web search for improving retrieval effectiveness,"2004.

[7] Y.Xu,K.Wang,B.Zhang and Z.Chen,"Privacy-Enhancing Personalized Web search,"Proc.16th Int'l Conf.,2007.

A Modified Model for Threat Assessment by Fuzzy Logic Approach

Ehsan Azimirad
PHD Student, Electrical and Computer Engineering
Department, Hakim Sabzevari University
Sabzevar, Iran

Javad Haddadnia
Associate Professor, Electrical and Computer Engineering
Department, Hakim Sabzevari University
Sabzevar, Iran

Abstract— **In this paper, a precise description of the threat evaluation process is presented. This is followed by a review describing which parameters that have been suggested for threat evaluation in an air surveillance context throughout the literature. Threat evaluation is a critical component of the system protecting the defended assets against the hostile targets like aircrafts, missiles, helicopters etc. The degree of threat is evaluated for all possible hostile targets on basis of heterogeneous parameter values extracted from various sensors, to improve the situational awareness and decision making. Taking into consideration the amount of uncertainty involved in the process of threat evaluation for dynamic targets, the fuzzy logic turns out to be a good candidate to model this problem. This model is based on a Fuzzy logic approach, making it possible to handle imperfect observations. The structure of the Fuzzy Logic is described in detail. Finally, an analysis of the system's performance as applied to a synthetic static scenario is presented. The simulation results are acceptable and fine and show that this model is reliability.**

Keywords-component; Threat Assessment; Fuzzy Knowledge Based System; Decision Support System; Weapons Assignment; Threat Evaluation Fuzzy Model

I. INTRODUCTION

In order to support the security of any nation the places of significance are to be protected as defended assets. The various defended assets can be air bases, tourist places, bridges, camps, nuclear power plants, command post, harbors, radars, monuments, parliament's buildings, etc. In the war as well as peace keeping scenario it becomes critical to understand the possible enemy dynamic targets such as aircrafts, missiles, helicopters, etc which can be manned or unmanned targets. In a military environment it is often the case that decision makers in real-time have to evaluate the tactical situation and to protect defended assets against enemy threats by assigning available weapon systems to them [1]. The decision making is very critical with respect to available resources and time. The dynamic targets are those targets which are mobile and exhibit change in their characteristic behavior. Various factors are considered for a decision making augmented with human cognitive intelligence. An expert system built with help of fuzzy logic can play an important role in enhancing situation awareness and automated decision making. The protection of defended assets is the prime objective of threat evaluation modeling of dynamic targets. An assumption is made that defending targets act as potential threats, but targets may be friend or enemy which is decided by IFF. The IFF is designed by command and control system. In this situation, prioritization of potential threats is very important according to threat level of detected enemy targets via multi resources. Battle space and intelligent sensors help in target classification. Threat value quantifies the possibility of threat or danger imposed by a potential target. In this situation of possible multiple targets, it becomes critical to prioritize the degree of threat involved with them to decide which target is more dangerous via predicting the threat value. Threat value is directly proportional to the amount of danger a target produces towards the protected asset. The higher threat value implies more dangerous target. This analysis in turn will play a significant role in weapon allocation against suspicious targets.

A grid of sensors produces large amount of heterogeneous data which can be used to evaluate the degree of threat of a target. Thus threat evaluation is a high level information fusion process. At times the threat evaluation becomes challenging in the presence of multiple parameters and processes. There is some amount of uncertainty involved in these parameters depending on the nature of targets and assets involved. It is difficult to formulate mathematical model by using selected parameters as inputs to generate the threat value as an output. The fuzzy inference system turns out to be one of the most efficient methods for the threat evaluation of dynamic targets under uncertain condition.

The remainder of this paper is organized as follows. In section II, a precise description of the threat evaluation process is presented. This is followed by a classification of parameters that throughout the literature are suggested for use in threat value calculation, and a summary of different algorithms and methods for threat evaluation that exist. In section III, a threat evaluation system based on the findings from the literature review is presented. The calculation of threat values in the system is performed by making inference in a fuzzy model. The structure of this fuzzy model is described, together with an analysis of the system's behavior as applied to a synthetic scenario. In section IV, a simulation and its results are presented. Finally, in section V the paper is concluded and thoughts regarding future work are presented.

II. THREAT DEFINITION

The threat is an expression of intension to inflict evil, injury, or damage [1, 12]. These threats are according to Steinberg [12] modeled in terms of relationships between threatening entities and threatened entities. The threatening entities will be referred to as targets, while the threatened entities are referred to as defended assets. The threat evaluation is significant component in target classification process. Small errors or mistakes in threat evaluation and target classification can result in huge damage of life and property.

A threat is often assessed as a combination of its capability and intent ([12], [14], [15]). A target's capability is its ability to inflict injury or damage to defended assets, while intent refers to its will or determination to inflict such damage [17]. In [15], a third threat component is mentioned: opportunity. This is spatio-temporal states of affairs making it possible to carry out one's intent given sufficient capabilities [16].

Threat evaluation helps in case of weapon assignment, and intelligence sensor support system. It is very important factor to analyze the behavior of enemy tactics as well as our surveillance. Disastrous situation in terms of loss of life and the valuable assets occur due to wrong evaluation of threat value. In this case we will suffer more as damages so it is important to evaluate more accurately.

Threat evaluation is a process based on defending targets to defended asset; here an assumption is to protect one asset against several defending targets but consideration of more number of assets will give realistic feel towards threat evaluation. It is a high level information fusion technique that belongs to third level data fusion model in Joint Directors of Laboratories (JDL) as seen in Fig. 1.

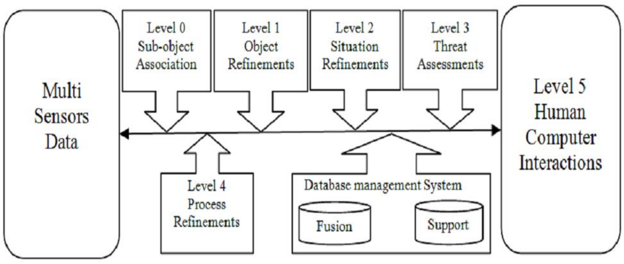

Multi Sensors Data
Level 0 Sub-object Association
Level 1 Object Refinements
Level 2 Situation Refinements
Level 3 Threat Assessments
Level 5 Human Computer Interactions
Level 4 Process Refinements
Database management System
Fusion
Support

Figure 1. The JDL Model

A. JDL Model

The JDL is a conceptual information fusion Model, which describes the processes, functions and specific techniques used for information fusion. Data fusion is the process of combining data or information to estimate or predict entity states [2]. It describes how data from different sources is transformed to information. This information used by decision makers which improves the situational awareness. In this model, data utilized is obtained from different sources like radars, sensors and databases. After estimation of information, aggregation and improvement can be done to extract right information for the decision makers. The JDL model comprises different levels [2]:

Level 0: Sub-Object Data Assessment: This level focuses on heterogeneous data collection. Assessment and prediction of data observable states on the basis of data association and characterization is done in this level. At this level, data is accessed from different sources, which may be localized or distributed. The main task of this level is to pre-process data by correcting biases and standardizing the input before the data from variety of sources is fused.

Level 1: Object assessment: The data collected in level zero is processed in this level to extract useful information. Assessment and prediction of entity states on the basis of observation-to track association for continuous state estimation and discrete state estimation is done in this level.

Level 2: Situation Assessment: The information extracted in level one is utilized to study the impact on current situation. Assessment and prediction of relations between the entities and relationship with the surrounding is focused in this level. This includes force structure, cross force relations, communications and etc.

Level 3: Threat Assessment: The situation information generated in level two is studied with respect to the role of possible contributors on the situation. Assessment and prediction of effects on situation of planned or estimated/predicted actions by the participants; to include interactions between action plans of multiple players is the main focus of this level.

Level 4: Process Refinement: This level focuses on the optimization of over all information fusion process that is an element of Resource Management. The Observe, Orient, Decide, and Act (OODA) loop is a concept originally applied to the combat operations process, often at the strategic level in military operations. The OODA loop is considered to be one of the most effective decisions making model in defense and security, often applied to understand commercial operations and learning processes today. The OODA loop can be seen in Fig. 2.

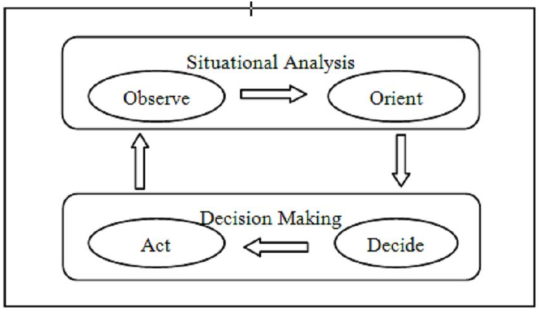


Figure 2. Situational Awareness OODA LOOP

The OODA loop is considered to be one of the most effective decisions making model in defense and security, often applied to understand commercial operations and learning processes today. When the enemy aircraft comes into radar contact, more direct information about the speed, size, and maneuverability of the enemy target becomes available. To determine which of several threats that represent the highest danger is of great importance, since errors such as prioritizing a lesser threat as a greater threat can result in engaging the wrong target, which often will have severe consequences [1].

III. THREAT EVALUATION MODEL

Consider a tactical situation where we have a set of defended assets $A = \{A_1, A_2, ..., A_m\}$ that we are interested in to protect (e.g. friendly forces, ships, bridges, and power plants). There is also a set of targets $T = \{T_1, T_2, ..., T_n\}$, which have been detected in the surveillance area. Now, the first problem is to for each target-defended asset pair (T_i, A_j), where $T_i \in T$ and $A_j \in A$, assign a threat value representing the degree of threat T_i poses to A_j, i.e., to define a function $Th(i, j): T \times A \rightarrow [0,1]$, assuming numbers between 0 and 1. Threat value of i th available defended asset from j th attacking target is $Th(i, j)$. The threat evaluation model is proposed in Fig. 3.

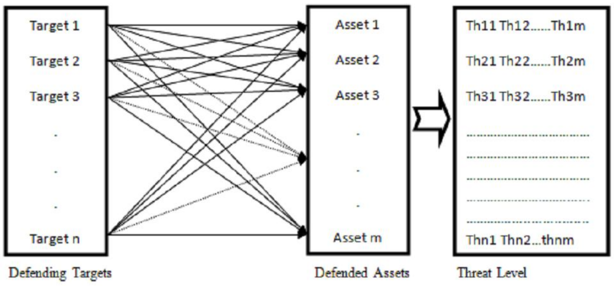


Figure 3. Asset- target pairs

The numbers 0 is lowest possible threat value and 1 is highest possible threat value.

A. Parameters for Threat Evaluation

In order to evaluate the threat posed by a target T_i on a defended asset A_j, there is a need to identify the parameters that control the threat value given a target-defended asset pair [20]. A large number of different parameters for threat value calculation have been suggested in the literature. However, many of these are closely related to each other.

The variety of parameters are proposed and used by researchers for threat evaluation [1]-[11]. These parameters have varying degree of effect on the threat value. Some parameters for calculating threat value are dependent on other parameters. A number of parameters [6] are discussed with their descriptions in Table I.

TABLE I. PARAMETERs TABLE

Attribute	Description
Speed	Approximate airspeed or an indication Of change (e.g., increasing).
Altitude	Approximate feet above ground or an indication of change (e.g., climbing).
Range/ Distance	The track's distance from own ship.
CPA (Closest Point of Approach)	Closest Point of Approach Estimated distance that track will pass by own ship if the track and own ship remain On their current courses.
Weapon envelope	The track's position with respect to its Estimated weapons envelope.
Own Support	Availability of nearby friendly ships Or patrol aircraft.
Visibility	Approximate number of miles, or an indication of atmospheric conditions (e.g., haze).
Maneuvers	Indicates the number of recent maneuvers, or if the track is following The ship.
Fire(Attack)	The Target Fire into Asset
IFF Mode	Identify Friend or Foe. Signals from a track that indicate if it is a friendly, or Perhaps neutral, aircraft.
Target Support	Availability targets for assistance to enemy target

Based on an exhaustive literature survey over publications dealing with threat evaluation in the information fusion domain and related areas, the parameters have been classified as follows.

1) Proximity parameters: An important class of parameters for assigning threat values to target-defended asset pairs is proximity parameters, i.e. CPA parameter. A key parameter that is used in many threat value evaluation techniques [1], [12] is the distance from the defended asset to the closest point of approach (CPA). The CPA is the point where the target eventually will be the closest to the defended asset, given current track estimates. Assuming stationary (non-moving) defended assets, this is the orthogonal projection of the position of the defended asset on the extension of the target's velocity vector (Fig. 4).

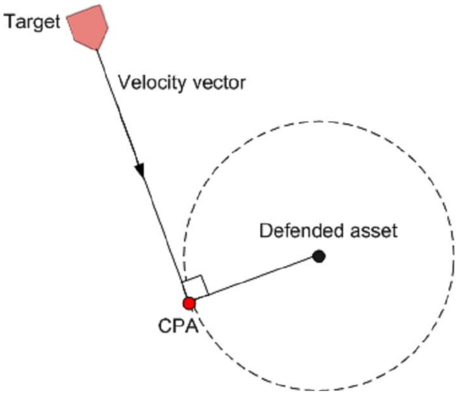


Figure 4. Closest point of approach between a target and a stationary Defended asset

The distance between the position of the defended asset and the CPA is clearly a measure of the threat level: the larger the distance, the less the threat. This distance will in the following be referred to as range from CPA.

2) Capability parameters: The next class of parameters for threat evaluation is capability parameters. This refers to the target's capability to threaten the defended asset. The several central parameters here are target type, weapon type and weapon envelope.

3) Intent parameters: The class of intent parameters is a broad category, containing parameters that can reveal something about the target's intent. The several parameters here are speed, heading [18], altitude and maneuvers [19].

B. *Design of a New Fuzzy Model*

In this paper one kind of rule-based algorithm is suggested, in which fuzzy inference rules are used to calculate the level of threat air targets pose to a navy combat ship, using speed, altitude, range, CPA, weapon envelope, own support, visibility, maneuver, fire, target support and IFF as input parameters and threat value as output parameter. For each input parameter, multiple membership functions are defined (e.g. slow, medium, and fast for the speed parameter and etc.). Such a membership function maps each point in the input space to a membership value between 0 and 1. Finally, fuzzy inference rules have been defined for how the input should affect the output parameter threat rating. The steps involved for threat value [21]:

1. Select target's information as inputs and threat rating as output. The target's information is collected from radar, and processed to set some information as fuzzy input evidences. The threat value rating is set as a fuzzy output.

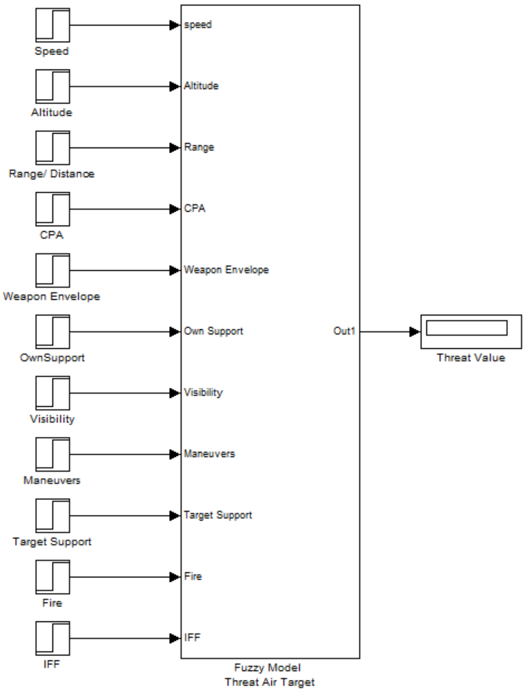


Figure 5. TEFM (Threat Evaluation Fuzzy Model)

2. Decide membership functions for each input and output parameters. Membership function of parameters is triangular. Membership functions of input parameters are in following Fig. 6 to Fig. 12.

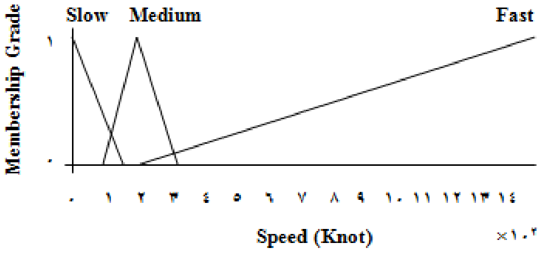


Figure 6. Membership functions for Speed

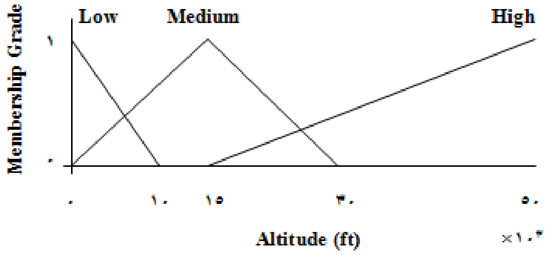


Figure 7. Membership functions for Altitude

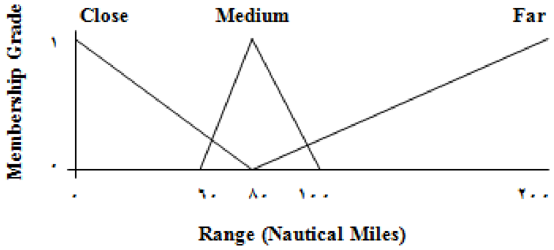


Figure 8. Membership functions for Range

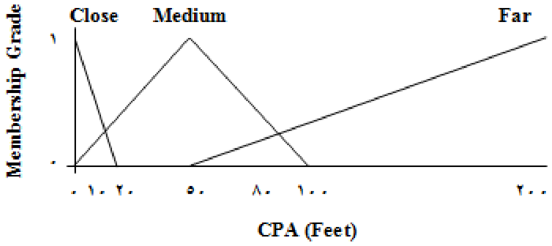


Figure 9. Membership functions for CPA

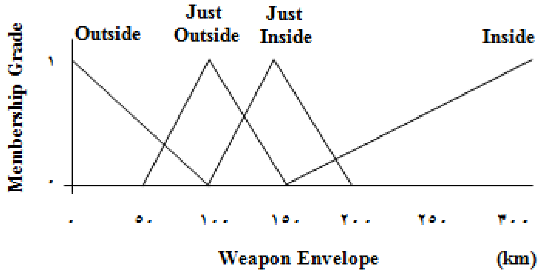


Figure 10. Membership functions for weapon envelope

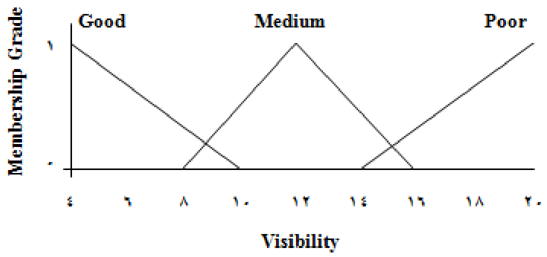
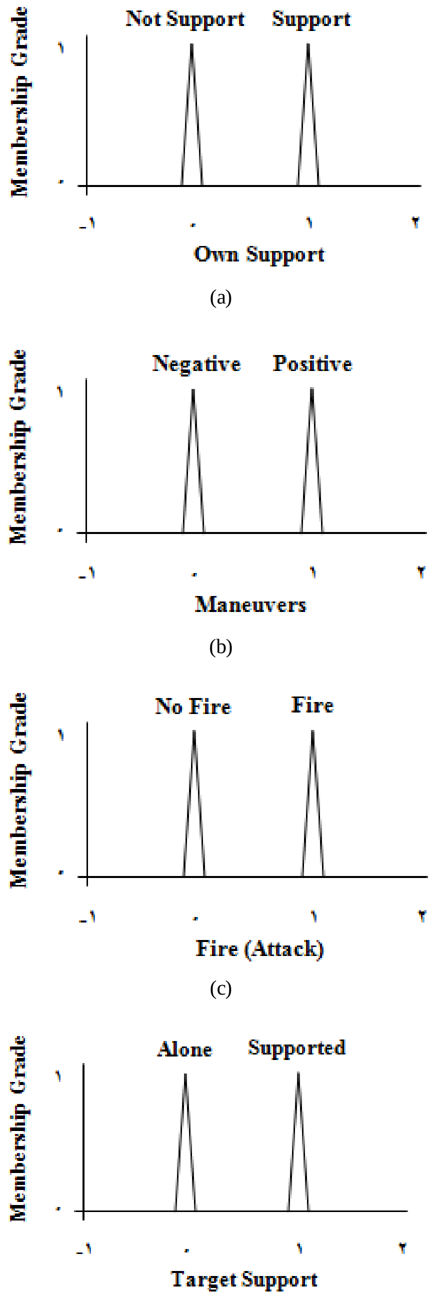


Figure 11. Membership functions for visibility

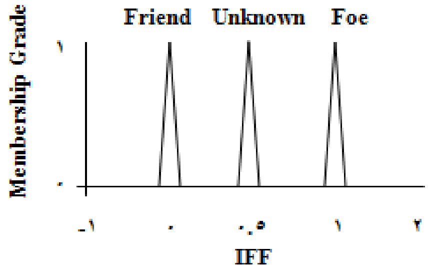


(a)

(b)

(c)

(d)



(e)

Figure 12. Membership functions for other parameters: (a) own support, (b) maneuvers, (c) fire, (d) target support, (e) IFF

Membership function of output parameter is the following Fig. 13.

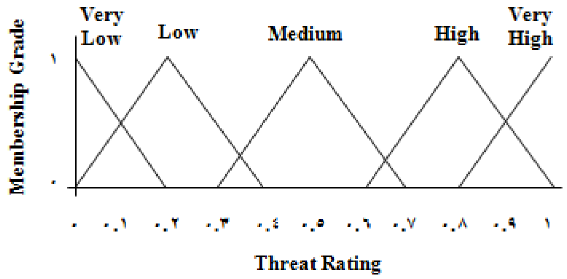


Figure 13. Membership functions for threat rating

3. Determine fuzzy rules by using inputs and output: Determine fuzzy inference rules using some standard data available and the expert's comments on the relation between the inputs Altitude, Speed, Range, CPA, weapon envelope, visibility, own support, maneuver, fire, target support and IFF and output Threat rating. Some tentative rules are framed and the results are evaluated for the validity of the results with respect to the real time and synthetic scenario. These inputs change the threat rating via rules. In this paper is defined 331 rules that has caused the system is robust and efficient. A few of fuzzy inference rules that have been used in the implementation are the following:

TABLE II. A FEW OF FUZZY INFERENCE RULES USED IN THIS PAPER

Rule Number	Description
Rule 1	IF (Altitude is low) AND (Speed is fast) AND (Range is close) AND (CPA is close) THEN (Threat Rating is very high) (Weight: 1).
Rule 2	IF (Altitude is high) AND (Speed is slow) AND (Range is far) AND (CPA is far) THEN (Threat Rating is very low) (Weight: 1).
Rule 3	IF (Altitude is medium) AND (Speed is medium) AND (Range is medium) AND (CPA is medium) THEN (Threat Rating is medium) (Weight: 1).
Rule 4	IF (Altitude is low) AND (Speed is fast) AND (Range is far) THEN (Threat Rating is medium) (Weight: 1).
Rule 5	IF (Altitude is high) AND (Speed is fast) AND (Range is far) THEN (Threat Rating is very low) (Weight: 1).

IV. THE RESULTS OF SIMULATION

To demonstrate the threat evaluation application, we have constructed a static test scenario. The scenario consists of a four defended asset and three air targets (one Boeing 747, one F-16, and one B-2 bomber).

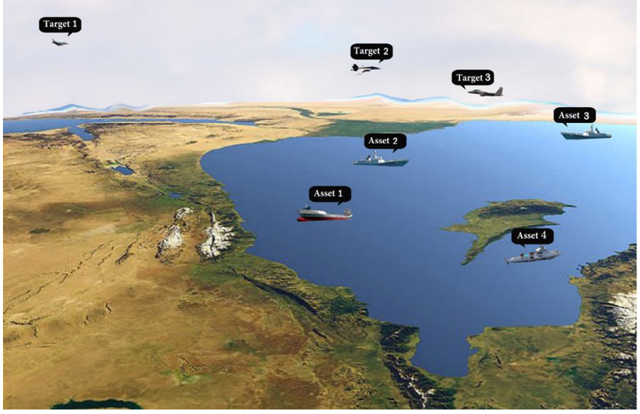


Figure 14. Projection of the targets used in the test scenario

A. Static Scenario

Simulation of block diagram proposed fuzzy model is completed for threat evaluation of targets by using the MATLAB software as seen in Figure 14. The figure shows a static scenario which is reads the input parameters as constant information in every time. This information is obtained from the radar system connected in the command and control unit. The underlying Fuzzy Inference System evaluates the value of threat for every one of the defended assets.

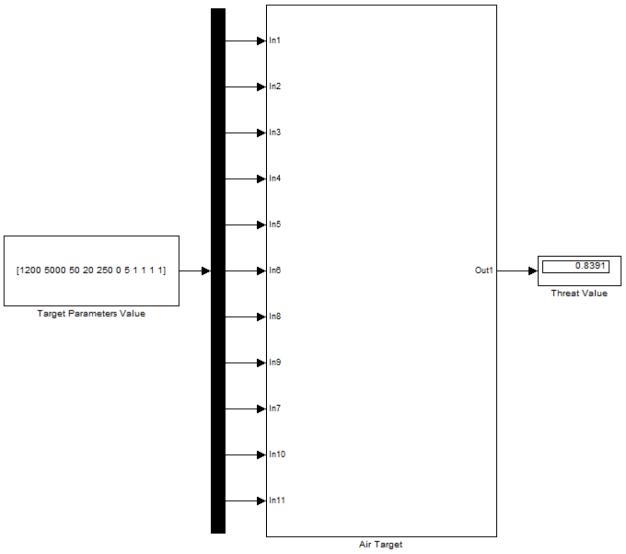


Figure 15. Static fuzzy model of threat evaluation in MATLAB

Simulation of this fuzzy model is done for the multiple set of inputs for the various example targets in static scenario. For example: For the input information like Altitude 5000 ft, Speed 1200 knot, range 50 in nautical miles, CPA 20 in ft, weapon envelope 250 km, visibility 5mA, own support 0 and other parameters are 1, the output generated is the threat rating 0.8391 (which lies between 0 and 1). It will change when values of parameter change time to time. Higher the threat rating identifies more dangerous target. The value of the threat rating will guide the decision making to engage the weapons in the process of protecting the assets from the targets. Simulation results in static test scenario for 25 instants are demonstrated in Table III and Table IV.

TABLE III. SIMULATION RESULTS IN 12 INSTANTS STATIC TEST SCENARIO

Static Scenarios	Speed	Altitude	Range	CPA	Weapon Envelope	Own Support	Visibility	Maneuver	Fire	Target Support	IFF	Threat Value
1	100	30000	180	80	50	1	5	0	0	0	0	0.1639
2	100	30000	180	80	50	0	5	0	0	0	0	0.2318
3	300	40000	180	150	50	1	5	0	0	0	0	0.1639
4	500	40000	180	150	50	1	5	0	0	0	0	0.1639
5	100	15000	120	80	50	0	5	0	0	0	0	0.1192
6	100	15000	120	80	50	1	5	1	1	0	0	0.5404
7	100	15000	120	80	50	1	5	1	1	1	1	0.6566
8	1200	5000	50	20	250	0	5	0	0	0	0	0.4636
9	1200	5000	50	20	250	1	5	0	0	0	0	0.3921
10	1200	5000	50	20	250	1	5	1	0	0	0	0.4872
11	1200	5000	50	20	250	1	5	1	1	0	0	0.8199
12	1200	5000	50	20	250	0	5	1	1	1	1	0.8391

TABLE IV. SIMULATION RESULTS IN OTHER 13 INSTANTS STATIC TEST SCENARIO

Static Scenarios	Speed	Altitude	Range	CPA	Weapon Envelope	Own Support	Visibility	Maneuver	Fire	Target Support	IFF	Threat Value
1	100	45000	160	40	150	0	10	0	0	0	0	0.1192
2	100	45000	160	40	150	0	10	1	1	0	0	0.5778
3	100	45000	160	40	150	1	10	1	1	0	0	0.5341
4	100	45000	160	40	150	1	10	1	1	1	1	0.6514
5	100	45000	160	40	150	0	10	1	1	1	1	0.6483
6	1000	10000	20	15	190	1	15	1	1	1	1	0.6483
7	1000	10000	20	15	190	0	15	1	1	1	1	0.6518
8	1000	10000	20	15	190	0	15	0	0	0	0	0.5158
9	1000	10000	20	15	190	1	15	0	0	0	0	0.4437
10	1000	10000	20	15	190	1	15	1	0	0	0	0.5436
11	1000	10000	20	15	190	1	15	1	1	0	0	0.7128
12	1000	10000	20	15	190	1	15	1	1	1	1	0.8335
13	1000	10000	20	15	190	0	15	1	1	1	1	0.8391

V. CONCLUSIONS

In this paper, we have given a precise description of the threat evaluation process. A literature review has been carried out regarding which parameters that have been suggested for threat value calculation throughout the literature, together with an overview of different algorithms that exist for threat evaluation. We have implemented a system for threat evaluation in an air defense environment. The underlying mechanism for threat evaluation in this system is a fuzzy logic.

The fuzzy logic based multi objective decision making system is an excellent tool available to deploy a decision support system. It simplifies the task of human decision maker to a great deal. Each target has different threat value at different time. In this paper, threat rating of targets is effectively estimated between 0 and 1 by using fuzzy inference system which is giving accurate result.

The implemented threat evaluation system has been applied to a synthetic air defense scenario. Also, for the first time in this paper, eleven parameters are introduced for threat evaluation such as altitude, speed, range, CPA, weapon envelope, own support, visibility, maneuver, fire, target support and IFF as input in fuzzy inference system. With regard to most parameters improved accuracy of threat assessment. An analysis of the system’s threat value calculations shows that the proposed fuzzy model works well for statically moving targets in fixed time.

The future work includes the design of fuzzy model is used to evaluate the threat of dynamically moving targets in real-time applications. Another interesting task is to investigate if the system’s calculated threat values on realistic scenarios agree with human experts on air defense.

The threat evaluation system can contribute significantly in the process of the improvement of situational awareness in peace and the battlefield scenarios in a network centric operation setup. This will add value to the battle space entity in a network centric platform operations with respect to the automated decision making support.

REFERENCES

[1] J. Roy, S. Paradis, M. Allouche, “Threat evaluation for impact assessment in situation analysis systems”, In: Kadar, I. (ed.) Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XI, vol. 4729, pp. 329–341, 2002.

[2] F. Johansson, “Evaluating the performance of TEWA System”, Orebro University, 2010.

[3] F. Johansson, G. Falkman, “A Bayesian network approach to threat evaluation with application to an air defense scenario,” *In: Proceedings of the 11th International Conference on Information Fusion*, 2008.

[4] T. Lampinen, J. Ropponen, T. T. Laitinen, “Joint Threat Assessment with Asset Profiling and Entity Bayes Net”, *In Proceeding of the 12th International Conference on Information Fusion*, Seattle, WA, USA, 2009.

[5] Y. Liang, “A fuzzy knowledge based system in situation and threat assessment’, *Journal of Systems Science & Information*, 4, 791–802, 2006.

[6] M. Liebhaber, B. Feher, “Air threat assessment: Research, model, and display guidelines”, *in Proceedings of the Command and Control Research and Technology Symposium*, 2002.

[7] Y. Liang, “An approximate reasoning model for situation and threat assessment”, in Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, 2007.

[8] X. Nguyen, “Threat assessment in tactical airborne environments”, in Proceedings of the Fifth International Conference on Information Fusion, 2002.

[9] S. Paradis, A. Benaskeur, M. Oxenham, P. Cutler, “Threat evaluation and weapons allocation in network-centric warfare”, *In: Proceedings of the 8th International Conference on Information Fusion*, 2005.

[10] T.J. Ross, Fuzzy Logic with Engineering Applications, Second Edition, John Wiley and Sons, 628.

[11] J.N. Roux, J.H. Van Vuuren, “Threat evaluation and weapon assignment decision support: A review of the state of the art, ORION”, vol. 23, pp. 151–186, 2007.

[12] J. Roy, S. Paradis, and M. Allouche, “Threat evaluation for impact assessment in situation analysis systems”, *in Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition* XI (I. Kadar, ed.), vol. 4729, pp. 329–341, July 2002.

[13] A. Steinberg, “An approach to threat assessment”, in Proceedings of the 8th International Conference on Information Fusion, 2005.

[14] X. Nguyen, “Threat assessment in tactical airborne environments”, in Proceedings of the Fifth International Conference on Information Fusion, 2002.

[15] E. L. Waltz and J. Llinas, “Multisensor Data Fusion”, Artech House, 1990.

[16] E. Little and G. Rogova, “An ontological analysis of threat and vulnerability,” *in Proceedings of the 9th International Conference on Information Fusion*, 2006.

[17] S. Paradis, A. Benaskeur, M. Oxenham, and P. Cutler, “Threat evaluation and weapons allocation in network-centric warfare,” *in Proceedings of the 8th International Conference on Information Fusion*, 2005.

[18] M. Oxenham, “Enhancing situation awareness for air defense via automated threat analysis,” *in Proceedings of the Sixth International Conference on Information Fusion*, vol. 2, pp. 1086–1093, 2003.

[19] M. Liebhaber and B. Feher, “Air threat assessment: Research, model, and display guidelines,” *in Proceedings of the 2002 Command and Control Research and Technology Symposium*, 2002.

[20] F. Johansson, G. Falkman, "A Bayesian network approach to threat evaluation with application to an air defense scenario", 11th International Conference on Information Fusion, 2008.

[21] S. Kumar, A. M. Dixit, “Threat Evaluation Modeling for Dynamic Targets Using Fuzzy Logic Approach”, *International Conference on Computer Science and Engineering*, 2012.

[22] VISIBILITY SENSOR Model 6000, THE STANDARD OF MEASUREMENT, Belfort Instrument Company, USA.

AUTHORS PROFILE

Ehsan Azimirad, received the B.Sc. degree in computer engineering and M.Sc. degree in control engineering with honors from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2006 and 2009, respectively. He is now PHD student in electrical and electronic engineering at Hakim Sabzevari University of Sabzevar in Iran. His research interests are fuzzy control systems and its applications in urban traffic, data fusion, threat assessment and any other problems, nonlinear control, Image Processing and Pattern Recognition and etc.

Javad Haddadnia, received his B.S. and M.S. degrees in electrical and electronic engineering with the first rank from Amirkabir University of Technology, Tehran, Iran, in 1993 and 1995, respectively. He received his Ph.D. degree in electrical engineering from Amirkabir University of Technology, Tehran, Iran in 2002. He joined Hakim Sabzevari University in Iran. His research interests include neural network, fuzzy logic and its applications in data fusion, threat assessment and applications, digital image processing, computer vision, and face detection and recognition. He has published several papers in these areas. He has served as a Visiting Research Scholar at the University of Windsor, Canada during 2001-2002. He is a member of SPIE, CIPPR, and IEICE.

A New Data Fusion Instrument for Threat Evaluation Using of Fuzzy Sets Theory

Ehsan Azimirad
PHD Student, Electrical and Computer Engineering
Department, Hakim Sabzevari University
Sabzevar, Iran

Javad Haddadnia
Associate Professor, Electrical and Computer Engineering
Department, Hakim Sabzevari University
Sabzevar, Iran

Abstract—**This paper represents an intelligent description of the threat evaluation process in 3 level of JDL model using of fuzzy sets theory. The degree of threat is evaluated for all possible targets to improve the situational awareness and decision making in command and control system and is calculated to precise weapon assignment. Taking into consideration the amount of uncertainty involved in the process of threat evaluation for dynamic targets, the fuzzy set theory turns out to be a good candidate to model this problem. In this approach, based on a fuzzy logic, is making it possible to handle imperfect observations. The structure of the fuzzy expert system based on fuzzy number approach is described in detail. Finally, an analysis of the system's performance as applied to multiple dynamic scenarios is presented. The simulation results show the correctness, accuracy, reliability and minimum errors in the system is designed.**

Keywords-component; Fuzzy Number, JDL Model, Decision Support System, Dynamic Air Targets, Multi Sensor Data Fusion; Weapons Assignment

I. Introduction

The main task of a battle management system, is integrating a floating sensor output data, target detection, diagnosis and management of a variety of weapons, targets and ultimately the decision is automatically installed on the vessel.

The heart of the battle management systems is data fusion system. Where the type and number of sensors and their reliability is very diverse and the sensors with heterogeneous output cannot overlap nor cannot independent or dependent. For increase the reliability of each of the parts of a battle management system use the fusion algorithms.

Data fusion is the software sector of battle management system. Data fusion is the process of combing information from a number of different sources to provide a robust and complete description of an environment or process of interest. Data fusion is of special significance in any application where a large amount of data must be combined, fused and distilled to obtain information of appropriate quality and integrity on which decisions can be made. Data fusion finds application in many military systems, in civilian surveillance and monitoring tasks, in process control and in information systems.

Data fusion methods are particularly important in the drive toward autonomous systems in all these applications. In principle, automated data fusion processes allow essential measurements and information to be combined to provide knowledge of sufficient richness and integrity that decisions may be formulated and executed autonomously [23].

Data fusion is often (somewhat arbitrarily) divided into a hierarchy of four processes. Levels 1 and 2 of this process are concerned with the formation of track, identity, or estimate information and the fusion of this information from several sources. Level 1 and 2 fusion is thus generally concerned with numerical information and numerical fusion methods (such as probability theory or kalman filtering). Level 3 and 4 of the data fusion process is concerned with the extraction of “ knowledge ” or decisional information. Very often this includes qualitative reporting or secondary sources of information or knowledge from human operators or other sources. Level 3 and 4 fusion is thus concerned with the extraction of high-level knowledge (situation awareness for example) from low level fusion processes, the incorporation of human judgment and the formulation of decisions and actions.

The various defended assets can be air bases, tourist places, bridges, camps, nuclear power plants, command post, harbors, radars, monuments, parliament’s buildings, etc. In the war as well as peace keeping scenario it becomes critical to understand the possible enemy dynamic targets such as aircrafts (bomber, fighter, and transporter), missiles, helicopters, etc which can be manned or unmanned targets.

In a military environment it is often the case that decision makers in real-time have to evaluate the tactical situation and to protect defended assets against enemy threats by assigning available weapon systems to them [11]. The dynamic targets are those targets which are mobile and exhibit change in their characteristic behavior. Various factors are considered for a decision making augmented with human cognitive intelligence.

An expert system built with help of fuzzy logic can play an important role in enhancing situation awareness and automated decision making. The protection of defended assets is the prime objective of threat evaluation modeling of dynamic targets. An assumption is made that defending targets act as potential threats, but targets may be friend or enemy which is decided by IFF (Identification, friend or foe). The IFF is designed by command and control system. In this situation, prioritization of potential threats is very important according to threat level (Degree of threat) of detected enemy targets via multi resources.

Battle space and intelligent sensors help in target classification. Threat value quantifies the possibility of threat or danger imposed by a potential target. In this situation of possible multiple targets, it becomes critical to prioritize the degree of threat involved with them to decide which target is more dangerous via predicting the threat value. Threat value is directly proportional to the amount of danger a target produces towards the protected asset. The higher threat value implies more dangerous target. This analysis in turn will play a significant role in weapon allocation against suspicious targets.

In a situation with several potential threats, it is of importance to prioritize these according to the degree of threat they represent to friendly defended assets, since such a degree indicates in which order the threats should be engaged [1], [21]. The degree of threat, known as threat value, can also be used to support intelligent sensor management [9], by allocating more sensor resources to targets with high threat values. To determine which of several threats that represent the highest danger is of great importance, since errors such as prioritizing a lesser threat as a greater threat can result in engaging the wrong target, which often will have severe consequences [12]. Threat evaluation is a high-level information fusion process that in relation to the JDL model of data fusion [22] belongs to level 3 [11], [9], [12], i.e. it is part of impact assessment.

A grid of sensors produces large amount of heterogeneous data which can be used to evaluate the degree of threat of a target. Thus threat evaluation is a high level information fusion process. At times the threat evaluation becomes challenging in the presence of multiple parameters and processes. There is some amount of uncertainty involved in these parameters depending on the nature of targets and assets involved. It is difficult to formulate mathematical model by using selected parameters as inputs to generate the threat value as an output. The fuzzy inference system turns out to be one of the most efficient methods for the threat evaluation of dynamic targets under uncertain condition. In this paper, we describe the importance of threat evaluation in introduction section.

Data fusion has its roots in the defense research community of the early 1980's. As a result the first data fusion models were either adapted from existing military oriented process models or were designed with a distinctly military flavor [24]. More recently the use of data fusion has broadened to include industrial, medical and commercial applications. More recent models have acknowledged this migration by reducing the military terminology. However, this still exists to some extent (and needs to be changed). Sensor network configuration, the display information and feedback within the network integration, some of the major issues in the implementation of a process model are considered. In Fig. 1, a data fusion model is presented for use in various applications. The model in this paper is useful JDL data fusion systems are usually discussed in the context of the military.

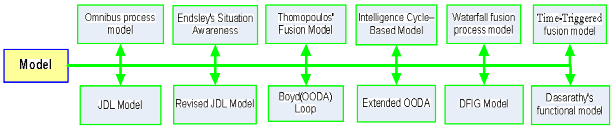

Model
Omnibus process model
Endsley's Situation Awareness
Thomopoulos' Fusion Model
Intelligence Cycle–Based Model
Waterfall fusion process model
Time-Triggered fusion model
JDL Model
Revised JDL Model
Boyd(OODA) Loop
Extended OODA
DFIG Model
Dasarathy's functional model

Figure 1. Multi sensor Data Fusion Models

The remainder of this paper is organized as follows. In section II, a precise description of the threat evaluation consisting of definition, modeling and evaluation in JDL model with threat parameters is presented. In section III fuzzy based approach for designing a new fuzzy model in fuzzy sets theory is presented. In section IV, simulation and results are presented. In this section, case study is demonstrated in four scenarios for static and dynamic targets. The calculation of threat values in the system is performed by making inference in a fuzzy model. The structure of this fuzzy model is described, together with an analysis of the system's behavior as applied to a synthetic scenario. Finally, in section V the paper is concluded and thoughts regarding future work are presented.

II. THREAT EVALUATION IN JDL MODEL

A. *Threat Definition*

The threat is an expression of intension to inflict evil, injury, or damage [1, 12]. These threats are ac-cording to Steinberg [12] modeled in terms of relationships between threatening entities and threatened entities. The threatening entities will be referred to as targets, while the threatened entities are referred to as defended assets. The threat evaluation is significant component in target classification process. Small errors or mistakes in threat evaluation and target classification can result in huge damage of life and property.

A threat is often assessed as a combination of its capability and intent ([12], [14], [15]). A target's capability is its ability to inflict injury or damage to defended assets, while intent refers to its will or de-termination to inflict such damage [17]. In [15], a third threat component is mentioned: opportunity. This is spatio-temporal states of affairs making it possible to carry out one's intent given sufficient capabilities [16].

Threat evaluation helps in case of weapon assignment, and intelligence sensor support system. It is very important factor to analyze the behavior of enemy tactics as well as our surveillance. Disastrous situation in terms of loss of life and the valuable as-sets occur due to wrong evaluation of threat value. In this case we will suffer more as damages so it is important to evaluate more accurately.

Threat evaluation is a process based on defending targets to defended asset; here an assumption is to protect one asset against several defending targets but consideration of more number of assets will give realistic feel towards threat evaluation. It is a high level information fusion technique that belongs to third level data fusion model in Joint Directors of Laboratories (JDL) as seen in Fig. 2.

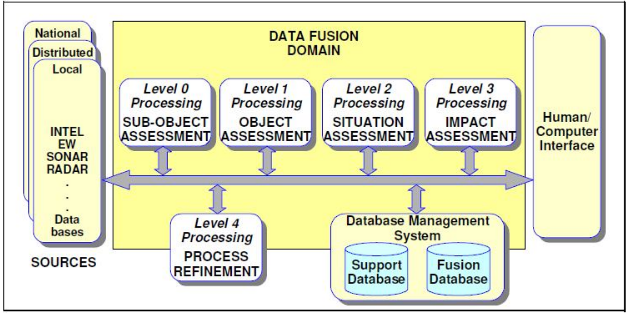


Figure 2. The JDL Model

B. JDL Model

The JDL is a conceptual information fusion Model, which describes the processes, functions and specific techniques used for information fusion. Data fusion is the process of combining data or information to estimate or predict entity states [2]. It describes how data from different sources is transformed to information. This information used by decision makers which improves the situational awareness. In this model, data utilized is obtained from different sources like radars, sensors and databases. After estimation of information, aggregation and improvement can be done to extract right information for the decision makers. The JDL model comprises different levels [2]:

Level 0: Sub-Object Data Assessment: This level focuses on heterogeneous data collection. Assessment and prediction of data observable states on the basis of data association and characterization is done in this level. At this level, data is accessed from different sources, which may be localized or distributed.

The main task of this level is to pre-process data by correcting biases and standardizing the input before the data from variety of sources is fused.

Level 1: Object assessment: The data collected in level zero is processed in this level to extract useful information. Assessment and prediction of entity states on the basis of observation-to track association for continuous state estimation (e.g. kinematics) and discrete state estimation (e.g. Target type and ID) is done in this level.

Level 2: Situation Assessment: The information extracted in level one is utilized to study the impact on current situation. Assessment and prediction of relations between the entities and relationship with the surrounding is focused in this level.

This includes force structure, cross force relations, communications, perceptual influences, physical context, etc.

Level 3: Threat Assessment: The situation information generated in level two is studied with respect to the role of possible contributors on the situation. Assessment and prediction of effects on situation of planned or estimated/predicted actions by the participants; to include interactions between action plans of multiple players (e.g. assessing susceptibilities and vulnerabilities to estimated threat actions given one's own planned actions) is the main focus of this level.

Level 4: Process Refinement: This level focuses on the optimization of over all information fusion process that is an element of Resource Management. The decision making in limited time is very important in most of the peace-keeping and war scenarios. The Observe, Orient, Decide, and Act (OODA) loop is a concept originally applied to the combat operations process, often at the strategic level in military operations. The OODA loop is considered to be one of the most effective decisions making model in defense and security, often applied to understand commercial operations and learning processes today. The OODA loop can be seen in Fig. 3 The whole idea of this loop is make faster decision for a quicker action with respect to the enemy.

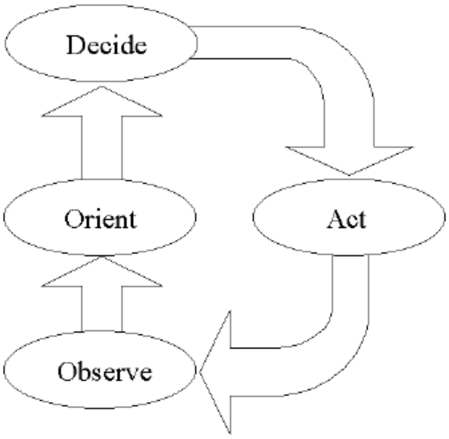


Figure 3. Situational Awareness OODA LOOP

The OODA loop is considered to be one of the most effective decisions making model in defense and security, often applied to understand commercial operations and learning processes today.

This Boyd control cycle or OODA loop represents the classic decision-support mechanism in military information operations. Because decision-support systems for situational awareness are tightly coupled with fusion systems, the Boyd loop has also been used for sensor fusion. Bedworth and O'Brien compared the stages of the Boyd loop to the JDL [33]:

- Observe – broadly comparable to the JDL level 0 and part of the collection phase of the intelligence cycle.
- Orient – encompasses the functions of JDL levels 1, 2 and 3. It also includes the structured elements of collection and the collation phases of the intelligence cycle.
- Decide – includes JDL level 4 (process refinement and resource management) and the dissemination activities of the intelligence community. It also includes much more (such as logistics and planning).

- Act – has no direct analogue in the JDL model and is the only model that explicitly closes the loop by taking account of the effect of decisions in the real world. The Fig. 4 represent OODA Loop and Mapping to JDL.

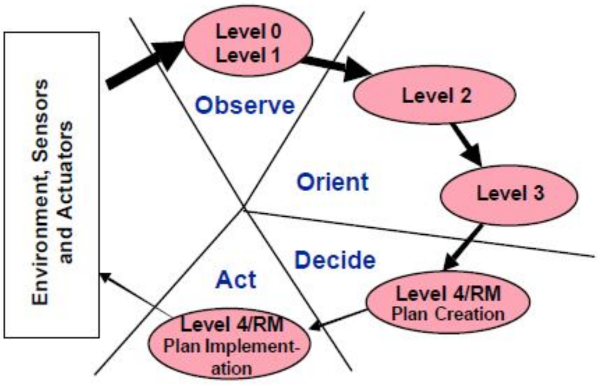


Figure 4. OODA Loop and Mapping to JDL

When the enemy aircraft comes into radar contact, more direct information about the speed, size, and maneuverability of the enemy target becomes available. To determine which of several threats that represent the highest danger is of great importance, since errors such as prioritizing a lesser threat as a greater threat can result in engaging the wrong target, which often will have severe consequences [1].

C. *Threat Modeling*

Consider a tactical situation where we have a set of defended assets $A = \{A_1, A_2, ..., A_m\}$ that we are interested in to protect (e.g. friendly forces, ships, bridges, and power plants). There is also a set of targets $T = \{T_1, T_2, ..., T_n\}$, which have been detected in the surveillance area. Now, the first problem is to for each target-defended asset pair (T_i, A_j), where $T_i \in T$ and $A_j \in A$, assign a threat value representing the degree of threat T_i poses to A_j, i.e., to define a function $Th(i,j): T \times A \to [0,1]$, assuming numbers between 0 and 1. Threat value of i th available defended asset from j th attacking target is $Th(i,j)$. The threat evaluation model is proposed in Fig. 5.

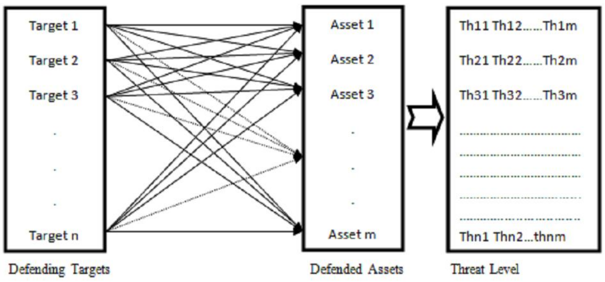


Figure 5. Asset- target pairs

The numbers 0 is lowest possible threat value and 1 is highest possible threat value.

D. *Parameters for Threat Assessment*

In order to evaluate the threat posed by a target T_i on a defended asset A_j, there is a need to identify the parameters that control the threat value given a target-defended asset pair [20]. A large number of different parameters for threat value calculation have been suggested in the literature. However, many of these are closely related to each other.

The variety of parameters are proposed and used by researchers for threat evaluation [1]-[11]. These parameters have varying degree of effect on the threat value. Some parameters for calculating threat value are dependent on other parameters. A number of parameters [6] are discussed with their descriptions in TABLE I.

TABLE I. PARAMETERs TABLE

Attribute	Description
Speed	Approximate airspeed or an indication Of change (e.g., increasing).
Altitude	Approximate feet above ground or an indication of change (e.g., climbing).
Range/ Distance	The track's distance from own ship.
CPA (Closest Point of Approach)	Closest Point of Approach Estimated distance that track will pass by own ship if the track and own ship remain On their current courses.
Weapon envelope	The track's position with respect to its Estimated weapons envelope.
Own Support	Availability of nearby friendly ships Or patrol aircraft.
Visibility	Approximate number of miles, or an indication of atmospheric conditions (e.g., haze).
Maneuvers	Indicates the number of recent maneuvers, or if the track is following The ship.
Fire(Attack)	The Target Fire into Asset
IFF Mode	Identify Friend or Foe. Signals from a track that indicate if it is a friendly, or Perhaps neutral, aircraft.
Target Support	Availability targets for assistance to enemy target

Based on an exhaustive literature survey over publications dealing with threat evaluation in the information fusion domain and related areas, the parameters have been classified as follows.

1) Proximity parameters: An important class of parameters for assigning threat values to target-defended asset pairs, i.e. CPA parameter.

2) Capability parameters: The next class of parameters for threat evaluation is capability parameters. This refers to the target's capability to threaten the defended asset. The several central parameters here are target type, weapon type and weapon envelope.

3) Intent parameters: The class of intent parameters is a broad category, containing parameters that can reveal something about the target's intent. The several parameters here are speed, heading [18], altitude and maneuvers [19].

III. FUZZY SETS THEORY

A. Fuzzy Inference

Fuzzy inference based on fuzzy sets theory is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns determined. The process of fuzzy inference involves all of the sections: Membership Function, Logical Operation, and If-Then Rules. Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, modeling & simulation, and computer vision. Because of its multidisciplinary nature, fuzzy inference systems are associated with a number of names, such as fuzzy rule- based systems, fuzzy expert systems, fuzzy modeling, fuzzy associative memory, fuzzy logic controllers, and simply fuzzy systems. There are two types of fuzzy inference system; Mamdani-type and Sugeno-type.

Mamdani's fuzzy inference method [5] is the most commonly used fuzzy methodology. Mamdani's method was among the first control systems built using fuzzy set theory. It was proposed in 1975 by Ebrahim Mamdani as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators. Mamdani's effort was based on Lotfi Zadeh's 1973 paper on fuzzy algorithms for complex systems and decision processes. Mamdani-type inference expects the output membership functions to be fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable that needs defuzzification. It is possible, and in many cases much more efficient, to use a single spike as the output membership functions rather than a distributed fuzzy set. This type of output is sometimes known as a singleton output membership function, and it can be thought of as a predefuzzified fuzzy set.

B. Design of a New Fuzzy Model

In this paper one kind of rule-based algorithm is suggested, in which fuzzy inference rules are used to calculate the level of threat air targets pose to a navy combat ship, using speed, altitude, range, CPA, weapon envelope, own support, visibility, maneuver, fire, target support and IFF as input parameters and threat value as output parameter. This matter is demonstrated in Fig. 6.

For each input parameter, multiple membership functions are defined. Such a membership function maps each point in the input space to a membership value between 0 and 1. Finally, fuzzy inference rules have been defined for how the input should affect the output parameter threat rating. The steps involved for threat value [21]:

1. Select target’s information as inputs and threat rating as output. Threat Evaluation Fuzzy Model is presented in Fig. 6.

2. Decide membership functions for each input and output parameters. Membership function of parameters is triangular. It is represented in mathematical expression:

$$\mu_{Speed} = \begin{cases} 0 \quad if \quad Speed \le a \\ \dfrac{(Speed - a)}{(b-a)} \quad if \quad a \le Speed \le b \\ \dfrac{(c - Speed)}{(c-b)} \quad if \quad b \le Speed \le c \\ 0 \quad if \quad Speed \ge c \end{cases} \tag{1}$$

Where a, b, and c stand for Starting point, Center point, and End point respectively.

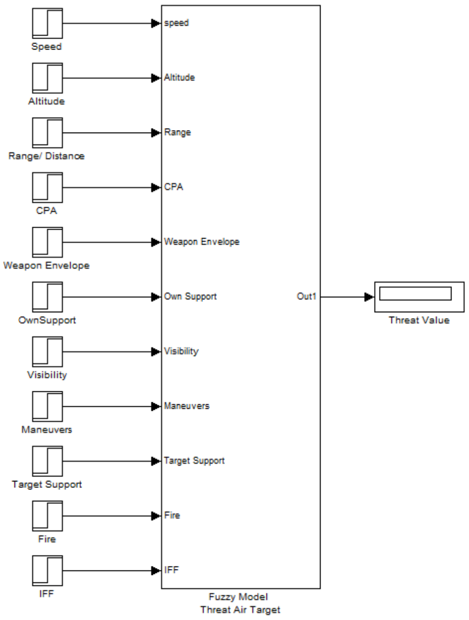


Figure 6. TEFM (Threat Evaluation Fuzzy Model)

Each of the parameters was explained in following: Generally Targets have maximum 1400 knot speed. E.g. Speed will lie in between 0 to 1400. The targets can achieve maximum 50000 ft Altitude but it depends on the type of target. Maximum range detected by the radar system will be 200 nautical miles but this range depends on the power of radar system. CPA can be calculated from velocity vector and position of asset. Maximum CPA is considered 200 feet. CPA will lie in between 0 to 200. Weapon envelope can be calculated as distance. Maximum weapon envelope is considered 300km for every of targets. Visibility can be calculated as current or voltage. Minimum and maximum visibility is considered between 4 to 20 mA [22]. These parameters are demonstrated in following.

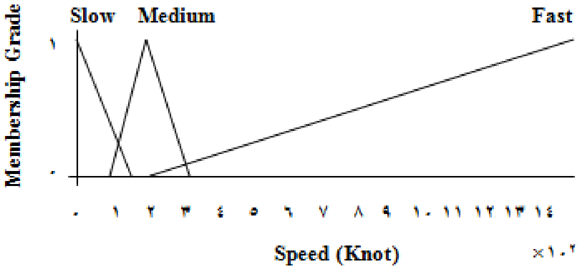


Figure 7. Membership functions for Speed

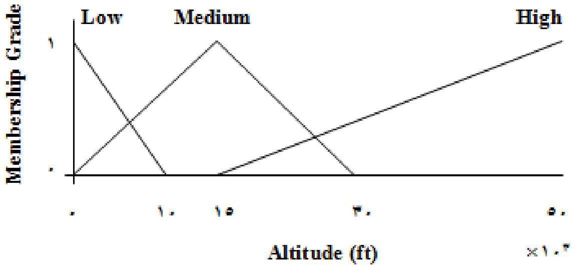


Figure 8. Membership functions for Altitude

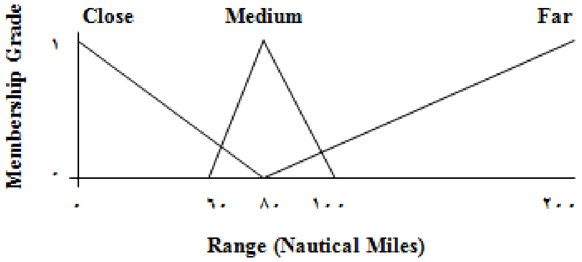


Figure 9. Membership functions for Range

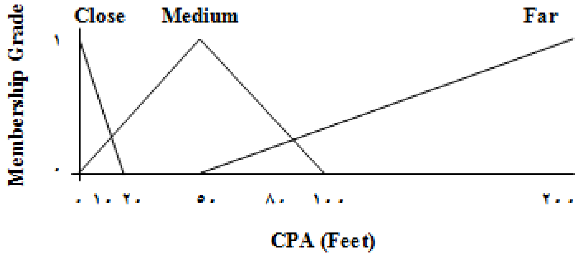


Figure 10. Membership functions for CPA

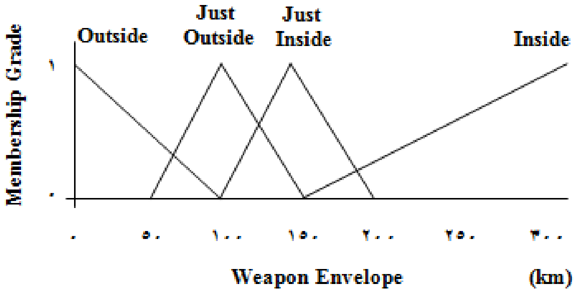


Figure 11. Membership functions for weapon envelope

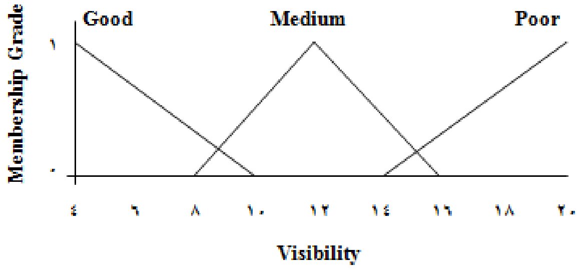
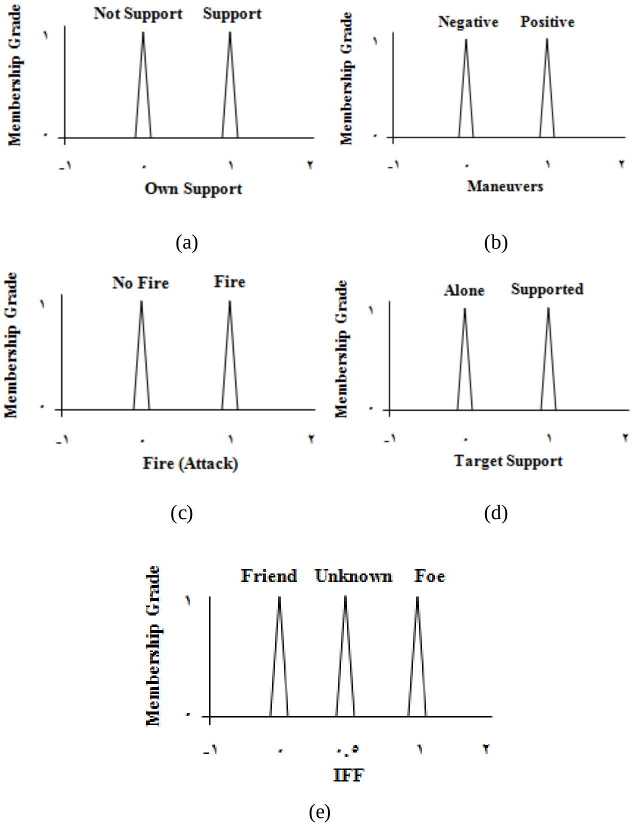


Figure 12. Membership functions for visibility

The residue of input parameters consists of own support, maneuver, fire, target support and IFF are considered 0 and 1 as seen in Fig. 13. These are singleton.



(a)

(b)

(c)

(d)

(e)

Figure 13. Membership functions for other parameters: (a) own support, (b) maneuvers, (c) fire, (d) target support, (e) IFF

The output parameter in threat evaluation of fuzzy model is threat rating that is between 0 and 1 as seen in Fig. 14.

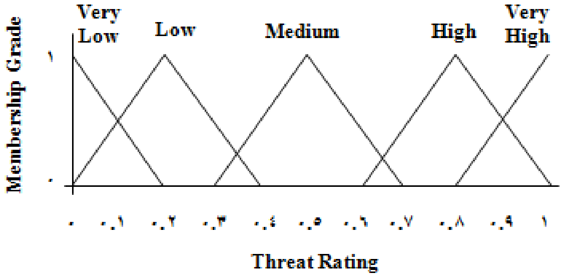


Figure 14. Membership functions for threat rating

3. Determine fuzzy rules by using inputs and output: Determine fuzzy inference rules using some standard data available and the expert's comments on the relation between the inputs and output. Some tentative rules are framed and the results are evaluated for the validity of the results with respect to the real time and synthetic scenario. These inputs change the threat rating via rules. In this paper is defined 331 rules that has caused the system is robust and efficient. A

few of fuzzy inference rules that have been used in the implementation are the following table:

TABLE II. A FEW OF FUZZY INFERENCE RULES USED IN THIS PAPER

Rule Number	Description
Rule 1	IF (Altitude is low) AND (Speed is fast) AND (Range is close) AND (CPA is close) THEN (Threat Rating is very high) (Weight: 1).
Rule 2	IF (Altitude is high) AND (Speed is slow) AND (Range is far) AND (CPA is far) THEN (Threat Rating is very low) (Weight: 1).
Rule 3	IF (Altitude is medium) AND (Speed is medium) AND (Range is medium) AND (CPA is medium) THEN (Threat Rating is medium) (Weight: 1).
Rule 4	IF (Altitude is low) AND (Speed is fast) AND (Range is far) THEN (Threat Rating is medium) (Weight: 1).
Rule 5	IF (Altitude is high) AND (Speed is fast) AND (Range is far) THEN (Threat Rating is very low) (Weight: 1).
Rule 6	IF (Altitude is low) AND (Speed is slow) AND (Range is close) AND (weapon envelope is outside) THEN (Threat Rating is very low) (Weight: 1).
Rule 7	IF (Altitude is low) AND (Speed is slow) AND (Range is close) AND (Weapon Envelope is inside) THEN (Threat Rating is high) (Weight: 1).
Rule 8	IF (Own Support is Not Support) AND (Fire (attack) is Fire) AND (Target Support is supported) AND (IFF is foe) THEN (Threat Rating is high) (Weight: 0.9).

IV. SIMULATION AND RESULTS

To demonstrate the threat evaluation application, we have constructed a test scenario. The scenario consists of a four defended asset and three air targets (one Boeing 747, one F-16, and one B-2 bomber). This scenario is discussed in four case studies for dynamic targets and is discussed in twelve samples for static targets of parameters. The Fig. 15 demonstrated battle environment.

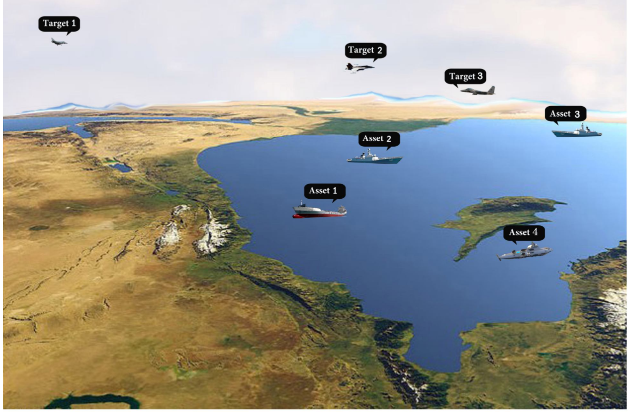


Figure 15. Combat environment in the test scenario

A. Static Scenario

Simulation of block diagram proposed fuzzy model is completed for threat evaluation of targets by using the MATLAB software as seen in Fig. 16. The figure shows a static scenario which is reads the input parameters as constant information in every time. This information is obtained from the radar system connected in the command and control unit. The underlying Fuzzy Inference System evaluates the value of threat for every one of the defended assets.

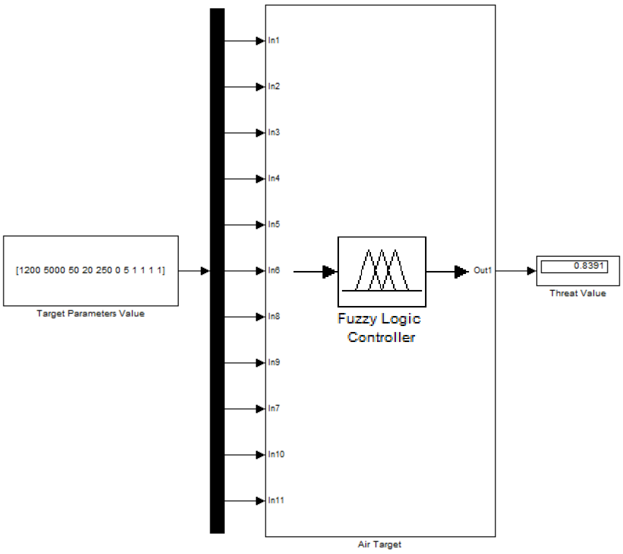
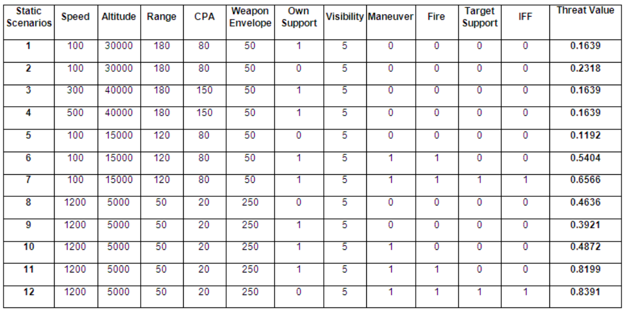


Figure 16. Static fuzzy model of threat evaluation in MATLAB

Simulation of this fuzzy model is done for the multiple set of inputs for the various example targets in static scenario. For example: For the input information like Altitude 5000 ft, Speed 1200 knot, range 50 in nautical miles, CPA 20 in ft, weapon envelope 250 km, visibility 5mA, own support 0 and other parameters are 1, the output generated is the threat rating 0.8391 (which lies between 0 and 1). It will change when values of parameter change time to time. Higher the threat rating identifies more dangerous target. The value of the threat rating will guide the decision making to engage the weapons in the process of protecting the assets from the targets. Simulation results in static test scenario for 12 instants are demonstrated in TABLE III.

TABLE III. SIMULATION RESULTS IN 12 INSTANTS STATIC TEST SCENARIO



Static Scenarios	Speed	Altitude	Range	CPA	Weapon Envelope	Own Support	Visibility	Maneuver	Fire	Target Support	IFF	Threat Value
1	100	30000	180	80	50	1	5	0	0	0	0	0.1639
2	100	30000	180	80	50	0	5	0	0	0	0	0.2318
3	300	40000	180	150	50	1	5	0	0	0	0	0.1639
4	500	40000	180	150	50	1	5	0	0	0	0	0.1639
5	100	15000	120	80	50	0	5	0	0	0	0	0.1192
6	100	15000	120	80	50	1	5	1	1	0	0	0.5404
7	100	15000	120	80	50	1	5	1	1	1	1	0.6566
8	1200	5000	50	20	250	0	5	0	0	0	0	0.4636
9	1200	5000	50	20	250	1	5	0	0	0	0	0.3921
10	1200	5000	50	20	250	1	5	1	0	0	0	0.4872
11	1200	5000	50	20	250	1	5	1	1	0	0	0.8199
12	1200	5000	50	20	250	0	5	1	1	1	1	0.8391

B. Dynamic Scenario

In this section, several scenarios for simulation dynamic air targets and threat evaluation them are discussed. Then for evaluating of robustness and efficiency of fuzzy model is done the comparison between them. The Fig. 17 shows a dynamic scenario which is reads the input parameters as information in real time problems that vary on time.

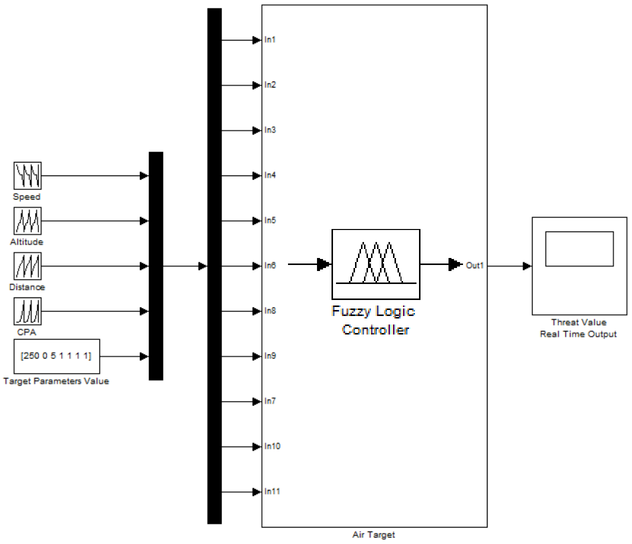


Figure 17. Dynamic fuzzy model of threat evaluation in MATLAB

B.1 The First Scenario

In the all of scenarios, inputs of fuzzy model consist of speed, altitude, range and CPA are stated on real time information and the other inputs are constant information (static) as seen in Fig. 17. In the first of scenario has been assumed that enemy target (Target1) be closed to defended asset (Asset1). The projection of combat environment used in the first of scenario is demonstrated in Fig. 18.

Figure 18. Projection of combat environment used in the one scenario

Every one of input parameters is varied as follow. The speed parameter of target with the increasing variable values is considered to be [153، 324، 564، 759، 900، 1040] knot in six different times. The altitude parameter of target with the decreasing variable values is considered to be [42830، 30025 ،21203، 17000، 10210، 5296] ft in six different times. The range and CPA parameters of target with the decreasing variable values respectively are considered to be [180، 143 ،125، 100، 80، 30] nautical miles and [173، 125، 90، 60، 40 ،10] feet in six different times. Also, weapon envelope is 250 km, visibility is 5mA, own support is 0 and the rest of the parameters are 1. The figure of time variation of each of the four parameters (speed, altitude, range and CPA) is shown as follows.

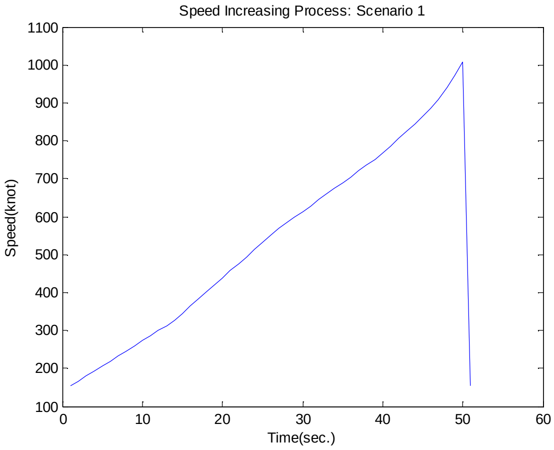


Figure 19. The time variation of the moving target speed

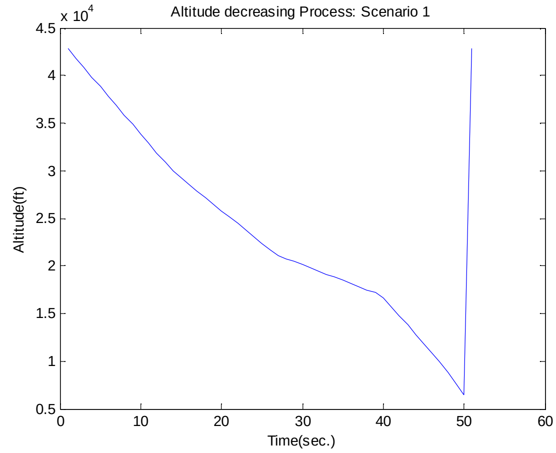


Figure 20. The time variation of the moving target altitude

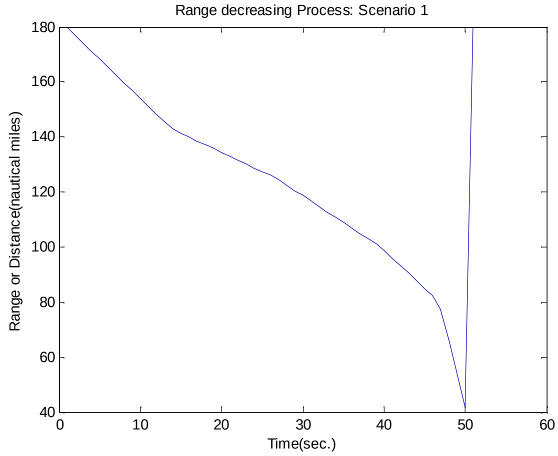


Figure 21. The time variation of the moving target range

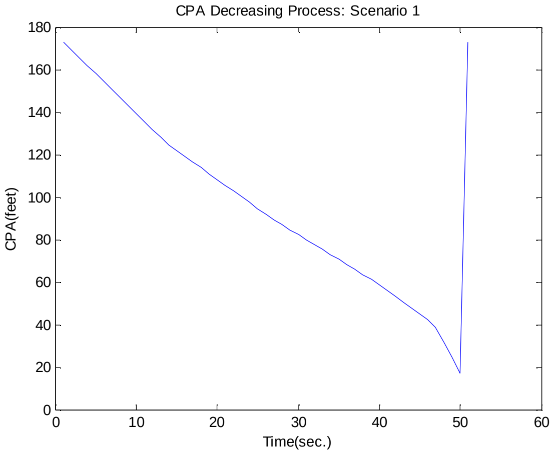


Figure 22. The time variation of the moving target CPA

The following figure shows the output of the fuzzy system threat assessment.

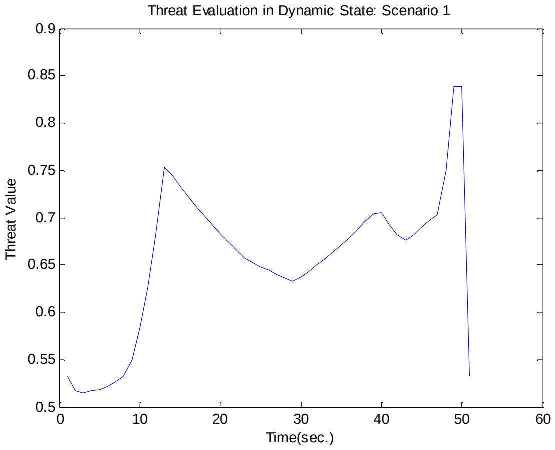


Figure 23. The output of threat fuzzy model in first scenario

By considering the amount of above, threat value increase significantly because the speed parameter is increased and three other parameters are decreased. The final threat value in the first scenario is got 0.8391. It shows that the threat is very high.

B.2 The Second Scenario

In the second of scenario has been assumed that enemy target (Target2) be fared from the defended asset (Asset2).

Every one of input parameters is varied as follow. The speed parameter of target with the decreasing variable values is considered to be [1300، 950، 764، 259، 190، 40] knot in six different times. The altitude parameter of target with the increasing variable values is considered to be [2000، 3153، 7000،10000، 15000، 19000] ft in six different times. The range and CPA parameters of target with the increasing variable values respectively are considered to be [15، 22، 36 ،78، 112، 140] nautical miles and [2، 5، 12، 15، 20، 35] feet in six different times. Also, weapon envelope is 150 km, visibility is 5mA, own support is 0 and the rest of the parameters are 0. The projection of combat environment used in the second of scenario is demonstrated in Fig. 24.

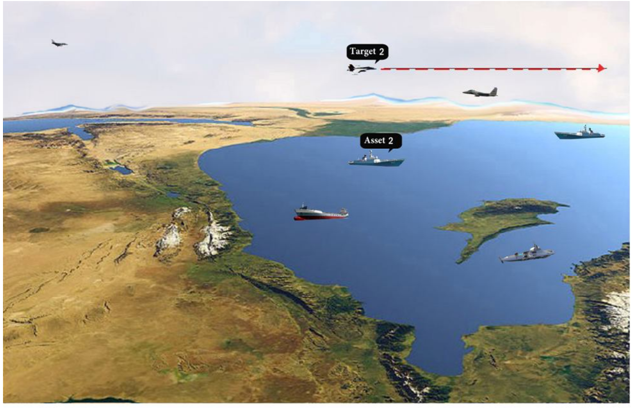


Figure 24. Projection of combat environment in the second scenario

The figure of time variation of each of the four parameters (speed, altitude, range and CPA) is shown as follows.

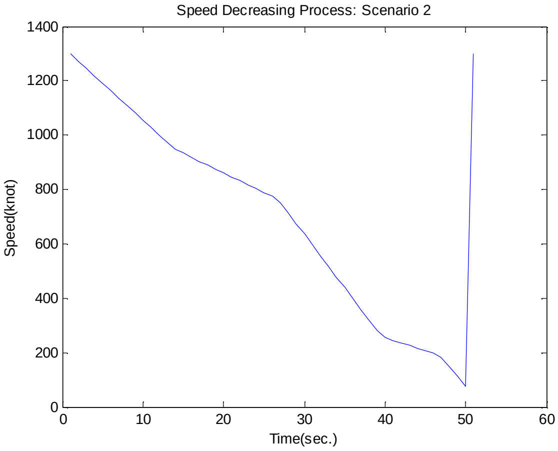


Figure 25. The time variation of the moving target speed

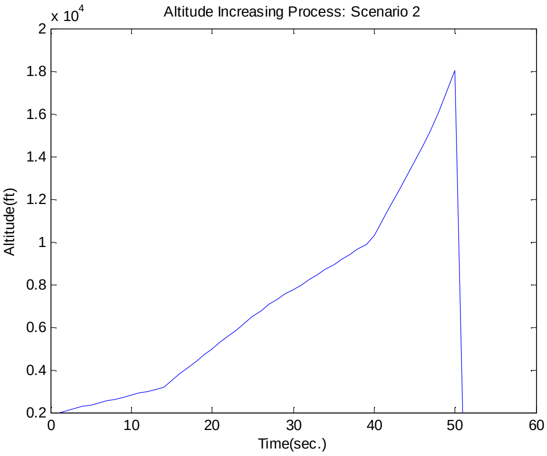


Figure 26. The time variation of the moving target altitude

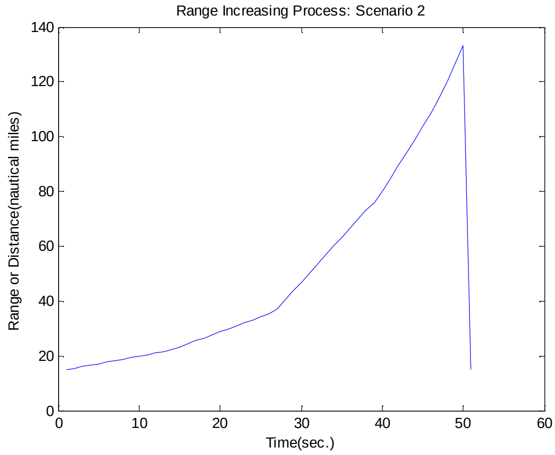


Figure 27. The time variation of the moving target range

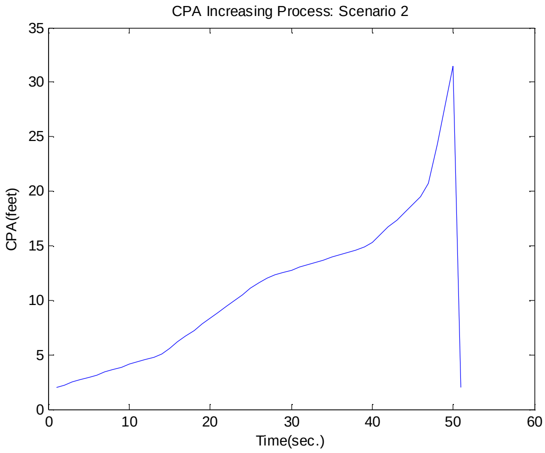


Figure 28. The time variation of the moving target CPA

The following figure shows the output of the fuzzy system threat assessment.

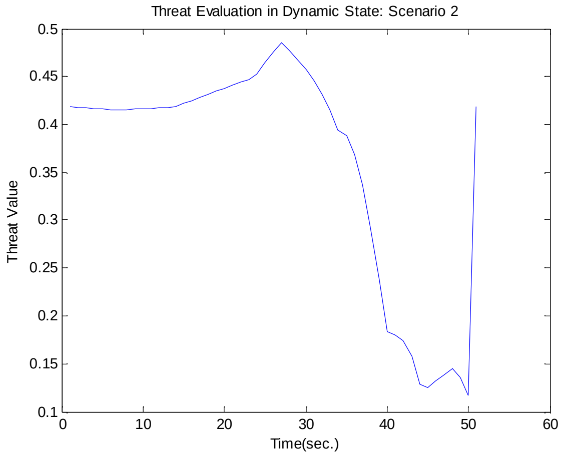


Figure 29. The output of threat fuzzy model in second scenario

By considering the amount of above, threat value decrease significantly because the speed parameter is decreased and three other parameters are increased. The final threat value in the second scenario is got 0.1164. It shows that the threat is very low.

B.3 The Third Scenario

In the third of scenario has been assumed that enemy target (Target3) the first be fared and then be closed to defended asset (Asset3). Every one of input parameters is varied as follow. The speed parameter of target with the increasing, decreasing and then increasing variable values is considered to be [250، 600، 764، 600، 890، 1150] knot in six different times. The altitude parameter of target with the increasing and then decreasing variable values is considered to be [3500، 8100، 15000، 12000 ،6000 ،1000] ft in six different times. The range and CPA parameters of target with the increasing and then decreasing variable values respectively are considered to be [15، 82، 166، 148، 122، 40] nautical miles and [160، 168، 190، 140، 90، 15] feet in six different times. Also, weapon envelope is 250 km, visibility is 5mA, own support is 0, maneuvers and fire is 1 and the rest of the parameters are 0. The projections of used in this scenario and the time variation of each of the four parameters are demonstrated in following figures.

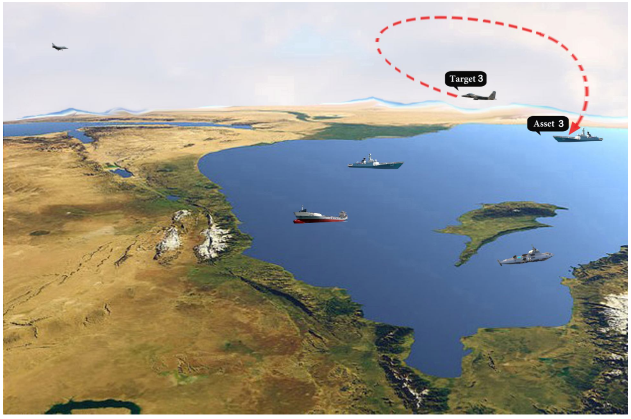


Figure 30. Projection of combat environment in the third scenario

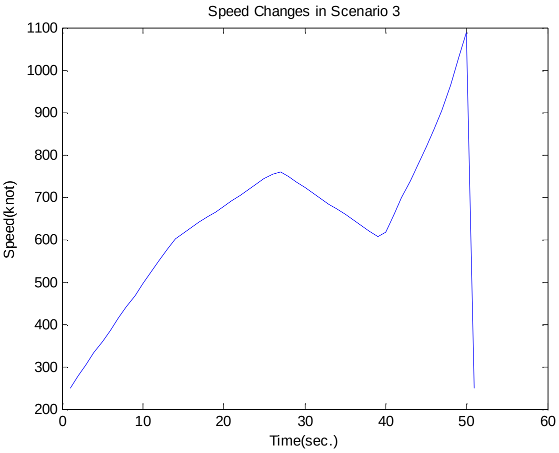


Figure 31. The time variation of the moving target speed

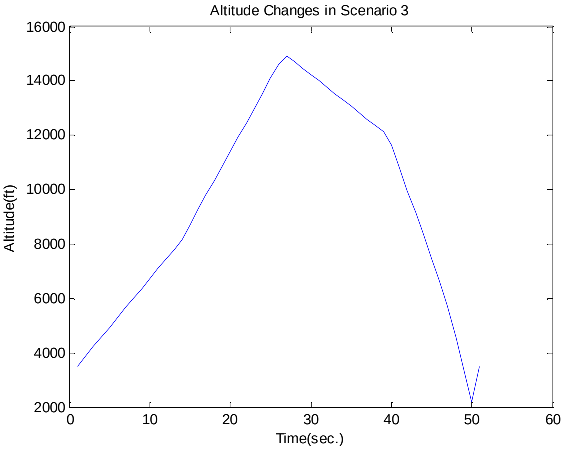


Figure 32. The time variation of the moving target altitude

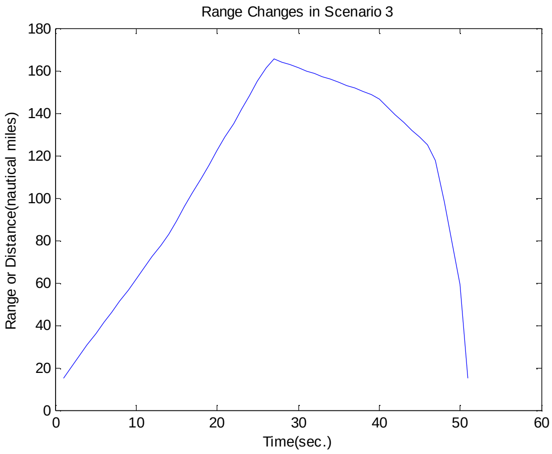


Figure 33. The time variation of the moving target range

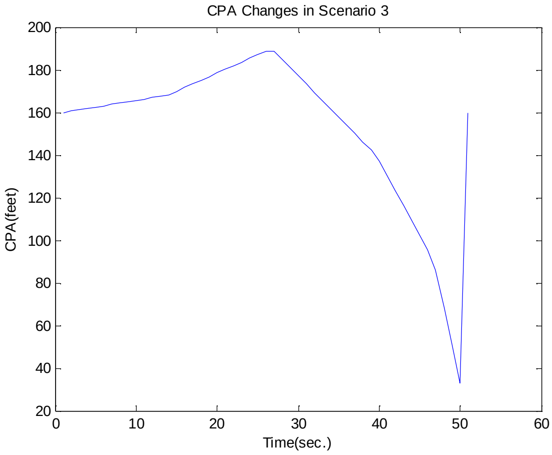


Figure 34. The time variation of the moving target CPA

The following figure shows the output of the fuzzy system threat assessment.

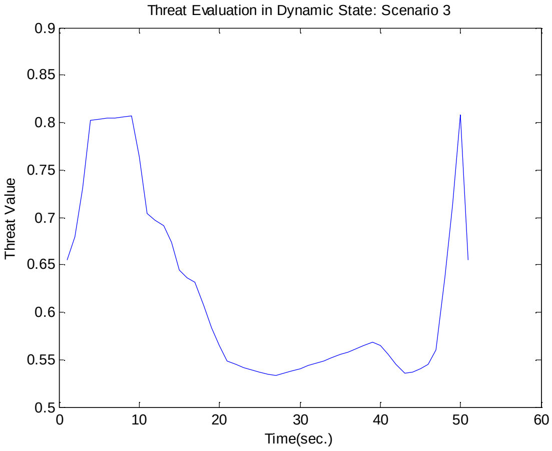


Figure 35. The output of threat fuzzy model in third scenario

By considering the amount of above, threat value increase significantly because of the speed parameter is increased and then decreased and increased. Also, three other parameters are increased and decreased. The final threat value in the third scenario is got 0.8087. It shows that the threat is very high.

B.4 The Four Scenario

In the four of scenario has been assumed that enemy target (Target4) the first be closed and then be fared from the defended asset (Asset4). Every one of input parameters is varied as follow. The speed parameter of target with the increasing and then decreasing variable values is considered to be [150، 350، 700، 1100، 1050، 800] knot in six different times. The altitude parameter of target with the decreasing and then increasing variable values is considered to be [35000، 25100، 15200، 9000 ،13000 ،20000] ft in six different times. The range and CPA parameters of target with the decreasing and then increasing variable values respectively are considered to be [180، 160، 100، 40، 80 ،170] nautical miles and [110، 80، 40، 10، 25، 35] feet in six different times. Also, weapon envelope is 100 km, visibility is 10mA and the rest of the parameters are 1. The projections of used in this scenario and the time variation of each of the four parameters are demonstrated in following figures.

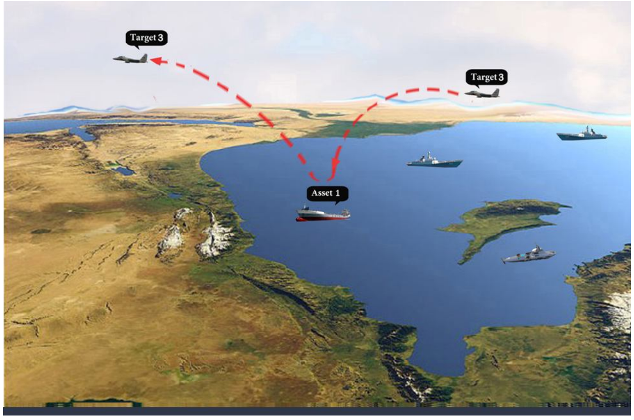


Figure 36. Projection of combat environment in the fourth scenario

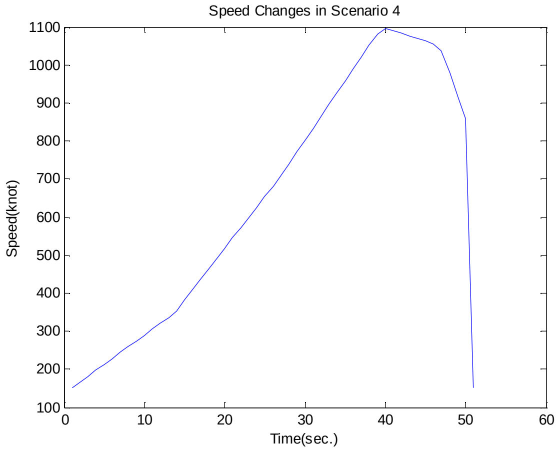


Figure 37. The time variation of the moving target speed

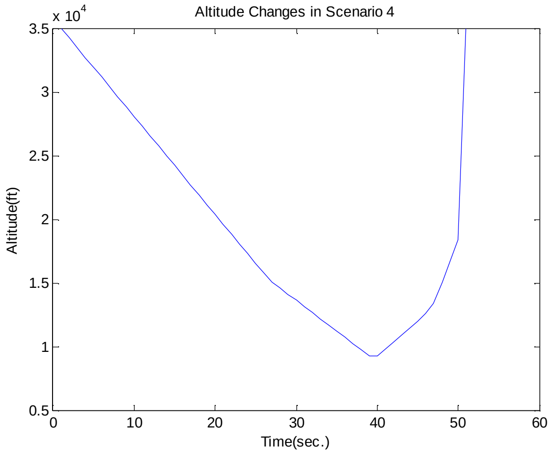


Figure 38. The time variation of the moving target altitude

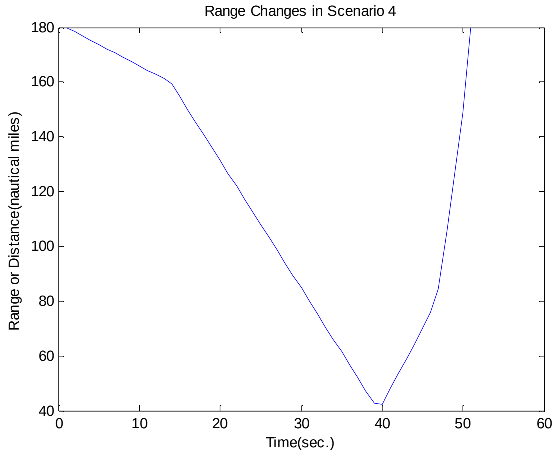


Figure 39. The time variation of the moving target range

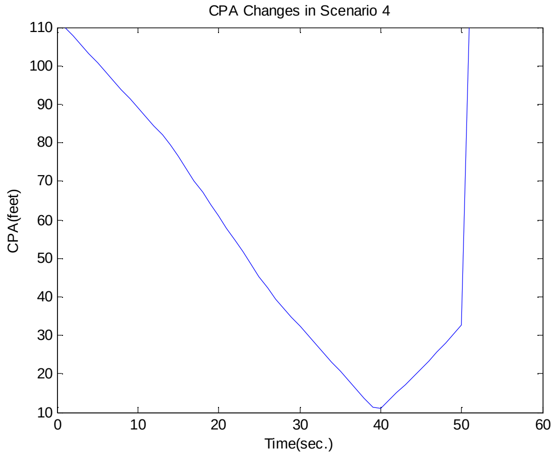


Figure 40. The time variation of the moving target CPA

The following figure shows the output of the fuzzy system threat assessment.

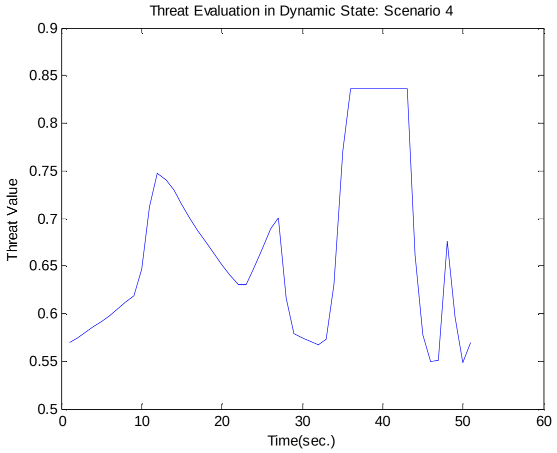


Figure 41. The output of threat fuzzy model in fourth scenarios

By considering the amount of above, threat value is averaged because of the speed parameter is increased and then decreased. Also, three other parameters are decreased and increased. The final threat value in the four scenarios is got 0.5483. It shows that the threat is medium.

Conclusion

In this paper, we have given a precise description of the threat evaluation process. A literature review has been carried out regarding which parameters that have been suggested for threat value calculation throughout the literature, together with an overview of different algorithms that exist for threat evaluation. We have implemented a system for threat evaluation in an air defense environment. The underlying mechanism for threat evaluation in this system is a fuzzy logic.

The fuzzy logic based multi objective decision making system is an excellent tool available to deploy a decision support system. It simplifies the task of human decision maker to a great deal. Each target has different threat value at different time. In this paper, threat rating of targets is effectively estimated between 0 and 1 by using fuzzy inference system which is giving accurate result.

The implemented threat evaluation system has been applied to a synthetic and time variant air defense scenario. Also, for the first time in this paper, eleven parameters are introduced for threat evaluation such as altitude, speed, range, CPA, weapon envelope, own support, visibility, maneuver, fire, target support and IFF as input in fuzzy inference system. With regard to most parameters improved accuracy of threat assessment. An analysis of the system's threat value calculations shows that the proposed fuzzy model works well for statically moving targets in fixed time.

The future work includes the design of fuzzy model is used to evaluate the threat of dynamically moving targets in real-time military applications. Another interesting task is to investigate if the system's calculated threat values on realistic scenarios agree with human experts on air defense.

The threat evaluation system can contribute significantly in the process of the improvement of situational awareness in peace and the battlefield scenarios in a network centric operation setup. This will add value to the battle space entity in a network centric platform operations with respect to the automated decision making support. Table IV is demonstrated the comparison of the results of four scenarios.

TABLE IV. THE FINAL THREAT VALUE IN FOUR SCENARIOS

scenario	Target Position	Final Threat Value	Output Type
1	Target be closed to Asset	0.8391	Very High
2	Target be fared the Asset	0.1164	Very Low
3	Target be fared and then closed to Asset	0.8087	Very High
4	Target be closed and then fared the Asset	0.5483	Medium

In this paper, an analysis of the system's performance as applied to multiple static and dynamic scenarios is presented. The simulation results in table 3 and 4 related to static targets are adapted with the results of table 5 related to dynamic targets in different scenarios. The results show the correctness, accuracy, reliability and minimum errors in the system is designed.

REFERENCES

[1] J. Roy, S. Paradis, M. Allouche, "Threat evaluation for impact assessment in situation analysis systems", In: Kadar, I. (ed.) Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XI, vol. 4729, pp. 329–341, 2002.

[2] F. Johansson, "Evaluating the performance of TEWA Sys-tem", Orebro University, 2010.

[3] F. Johansson, G. Falkman, "A Bayesian network approach to threat evaluation with application to an air defense scenario", In: Proceedings of the 11th International Conference on Information Fusion, 2008.

[4] T. Lampinen, J. Ropponen, T. T. Laitinen, "Joint Threat Assessment with Asset Profiling and Entity Bayes Net", In Proceeding of the 12th International Conference on Information Fusion, Seattle, WA, USA, 2009.

[5] Y. Liang, "A fuzzy knowledge based system in situation and threat assessment", Journal of Systems Science & Information, 4, 791–802, 2006.

[6] M. Liebhaber, B. Feher, "Air threat assessment: Research, model, and display guidelines", in Proceedings of the Command and Control Research and Technology Symposium, 2002.

[7] Y. Liang, "An approximate reasoning model for situation and threat assessment", in Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, 2007.

[8] X. Nguyen, "Threat assessment in tactical airborne environ-ments", in Proceedings of the Fifth International Conference on Information Fusion, 2002.

[9] S. Paradis, A. Benaskeur, M. Oxenham, P. Cutler, "Threat evaluation and weapons allocation in network-centric warfare", In: Proceedings of the 8th International Conference on Information Fusion, 2005.

[10] T. J. Ross, "Fuzzy Logic with Engineering Applications", Second Edition, John Wiley and Sons, 628.

[11] J. N. Roux, J. H. Van Vuuren, "Threat evaluation and wea-pon assignment decision support: A review of the state of the art, Orion", vol. 23, pp. 151–186, 2007.

[12] J. Roy, S. Paradis, M. Allouche, "Threat evaluation for impact assessment in situation analysis systems", in Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XI (I. Kadar, ed.), vol. 4729, pp. 329–341, July 2002.

[13] A. Steinberg, "An approach to threat assessment", in Proceedings of the 8th International Conference on Information Fusion, 2005.

[14] X. Nguyen, "Threat assessment in tactical airborne environ-ments", in Proceedings of the Fifth International Conference on Information Fusion, 2002.

[15] E. L. Waltz, J. Llinas, "Multisensor Data Fusion", Artech House, 1990.

[16] E. Little, G. Rogova, "An ontological analysis of threat and vulnerability", in Proceedings of the 9th International Conference on Information Fusion, 2006.

[17] S. Paradis, A. Benaskeur, M. Oxenham, P. Cutler, "Threat evaluation and weapons allocation in network-centric warfare", in Proceedings of the 8th International Conference on Information Fusion, 2005.

[18] M. Oxenham, "Enhancing situation awareness for air defence via automated threat analysis", in Proceedings of the Sixth International Conference on Information Fusion, vol. 2, pp. 1086–1093, 2003.

[19] M. Liebhaber, B. Feher, "Air threat assessment: Research, model, and display guidelines", in Proceedings of the 2002 Command and Control Research and Technology Symposium, 2002.

[20] F. Johansson, G. Falkman, "A Bayesian network approach to threat evaluation with application to an air defense scenario", 11th International Conference on Information Fusion, 2008.

[21] S. Kumar, A. M. Dixit, "Threat Evaluation Modelling for Dynamic Targets Using Fuzzy Logic Approach", International Conference on Computer Science and Engineering, 2012.

[22] Visibility Sensor Model 6000, "The Standard of Measure-ment", Belfort Instrument Company, USA.

[23] H. Durrant-Whyte, "Multi Sensor Data Fusion", Australian Centre for Field Robotics the University of Sydney, January 2001.

[24] P. Valin, E. Bosse, A. Jouan, "Airborne application of information fusion algorithms to classification", Technical Report TR 2004-282, Defense Research and Development Canada – Valcartier, May 2006.

[25] F. White, "A Model for Data Fusion", Proceedings 1st National Symposium on Sensor Fusion, 1988.

[26] A. Steinberg, C. Bowman, F. White, "Revisions to the JDL Data Fusion Model", SPIE, Vol. 3719, pp. 430-441, 1999.

[27] L. Klein, "Sensor and Data Fusion Concepts and Applica-tions", SPIE Volume TT14, 1993.

[28] S.J. Yang, J. Holsopple, M. Sudit, "Evaluating Threat Assess-ment for Multi-Stage Cyber Attacks", In Proceedings of the 2006 Military Communications Conference, Washington, DC. Oct. 23-25, 2006.

[29] R. Chinchani, A. Iyer, H.Q. Ngo, S. Upadhyaya, "Towards a theory of insider threat assessment", In Proceedings of the 2005 International Conference on Dependable Systems and Networks, 2005.

[30] F. BOLDERHEIJ, PHD thesis, "Mission Driven" Netherlands Defense Academy and the Centre for Automation of Mission Critical Systems (CAMS) – Force Vision. 2007.

[31] G. A. McIntyre, K. J. Hintz, "A Comprehensive Approach to Sensor Management, Part I: A Survey of Modern Sensor Management Systems", IEEE Transactions on SMC, April 1999.

[32] A. Erhard, S. McGalliard, "Advances in Military Multi- Sensor Data Fusion Techniques and Applications for Civilian Use", Eleventh Annual Freshman Conference, April 9, 2011.

[33] M. Bedworth, J. O'Brien, "The Omnibus Model: A New Model of Data Fusion?", 1999.

AUTHORS PROFILE

Ehsan Azimirad, received the B.Sc. degree in computer engineering and M.Sc. degree in control engineering with honors from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2006 and 2009, respectively. He is now PHD student in electrical and electronic engineering at Hakim Sabzevari University of Sabzevar in Iran. His research interests are fuzzy control systems and its applications in urban traffic, data fusion, threat assessment and any other problems, nonlinear control, Image Processing and Pattern Recognition and etc.

Javad Haddadnia, received his B.S. and M.S. degrees in electrical and electronic engineering with the first rank from Amirkabir University of Technology, Tehran, Iran, in 1993 and 1995, respectively. He received his Ph.D. degree in electrical engineering from Amirkabir University of Technology, Tehran, Iran in 2002. He joined Hakim Sabzevari University in Iran. His research interests include neural network, fuzzy logic and its applications in data fusion, threat assessment and applications, digital image processing, computer vision, and face detection and recognition. He has published several papers in these areas. He has served as a Visiting Research Scholar at the University of Windsor, Canada during 2001-2002. He is a member of SPIE, CIPPR, and IEICE.

Optimizing TCP Vegas for Optical Networks: a Fuzzy Logic Approach

Reza Poorzare[1], **Shahram Jamali**[2]

[1] *Department of Computer Science*
Young Researchers Club, Ardabil Branch, Islamic Azad University, Ardabil, Iran

[2] *Department of Computer Engineering*
University of Mohaghegh Ardabili, Ardebil, Iran

***Abstract*—Performance of TCP is reduced over buffer-less optical burst switched (OBS) networks by misunderstanding of the congestion status in the network. In other words, when a burst drop occurs in the network and we cannot distinguish congestion and burst contention in the network, TCP wrongly decreases the congestion window size (cwnd) and causes significant reduction of the network performance. This paper employs the fuzzy logic to solve this problem. By using the fuzzy logic we provide a framework to distinguish whether the burst drop is due to the congestion or is due to the burst contention. The full approach, for detecting state of network, relies on Round-Trip-Time (RTT) measurement only. So, this is an end-to-end scheme which only end nodes are needed to cooperate. Extensive simulative studies show that the proposed algorithm outperforms other TCP flavors such as TCP Vegas, TCP Sack and TCP Reno, in terms of throughput, packet delivery count and fairness.**

Key Words — Fuzzy Logic, Optical Burst Switching, TCP Vegas, Transport Control Protocol (TCP).

I. Introduction

Researchers have conducted tremendous amount of studies on TCP for customizing it to new networks [1-8]. These studies are categorized to three groups: Loss-based TCP (such as TCP Reno [5] and TCP Sack [6]), delay-based TCP (such as TCP Vegas [7] and Fast TCP [2,3]) and explicit notification-based TCP (such as XCP [8]). One of these new networks is optical burst switching (OBS) networks. OBS is a switching technique in context of wavelength division multiplexing (WDM) optical network to deal with large amount of Internet traffic [9]. In this network packets that arrived to the edge nodes are aggregated to bursts and these bursts are sent through the optical backbone. It means that we have two kinds of nodes in the networks, i.e. edge nodes and core nodes. Edge nodes assemble arrived packets to make burst and disassemble received bursts. Ingress nodes assemble packets to burst and egress nodes disassemble bursts to packets. Core nodes forward bursts in the optical network.

In OBS networks before sending bursts, edge node sends a control packet to reserve resources in the core nodes, in a process called burst reservation. This control packet is sent over an out-of-band channel. The control packets contain information about the burst such as the burst length and offset time. Offset time is the delay time between sending control packet and its corresponded burst [10]. Bandwidth in a fiber channel is enough to carry a large amount of traffic. Because of that in OBS network a fiber link is divided to several distinct channels to carry burst separately, this is called wavelength division multiplexing (WDM).

One of the main problems in OBS networks is burst contention. It means there is a probability that bursts have contention in the network, and it causes burst drops. Naturally, TCP may misinterpret source of this burst drop and wrongly suppose that it is due to the congestion.

Finding a solution to separate burst drop caused by congestion and burst drop caused by contention is an important consideration in OBS networks. One of the solutions to cope with the TCP false congestion detection problem is explicit signaling from the OBS layer to the TCP layer that is proposed in [11]. But this approach has some shortcomings. Generating explicit signal for every random burst contention increases network overhead, so performance of the network is decreased. Other approaches to solve false congestion detection problem in OBS networks include burst retransmission and deflection scheme at the OBS layers [12-14]. By using this scheme we can hide some of the bursts loss events from the upper TCP layer, therefore we reduce chance of false congestion detection problem in the network. With burst retransmission or deflection, contented bursts are retransmitted at the edge nodes or can be deflected to alternative routes, respectively. If deflection routing is enabled in the network, we have to use optical buffers in the networks and this is the drawback of the scheme. A threshold-based TCP Vegas is proposed in [15]. This scheme adjusts size of congestion window based on round trip times (RTTs) of packets received at TCP senders. If the number of RTTs that are longer than minimum RTTs exceeds the threshold, it means congestion happens in the network, otherwise there is no congestion in the network. To prevent data loss during random contention in the optical core, coordinated burst cloning and forward segment redundancy has been proposed in [9]. In this scheme, redundant segments are appended to each burst at the edge nodes and redundant burst segmentation (RBS) is implemented in the cores, so that when a contention occurs, primarily redundant data is dropped. The drawback of this scheme is appending redundant to the bursts increase overhead of the network. A study about impact of TCP synchronization of capacity dimensioning of Optical Burst Switched links has been done in [16]. When a burst with segment from different TCP flow is discarded, the synchronization of TCP flows appears and it causes all TCP flows reduce transmission rate at the same time. This paper analyzes the bandwidth capacity that needs to be provisioned in OBS links transporting synchronized flows when compared to a non-synchronized scenario. Three different variants of TCP that are TCP Tahoe, TCP Reno and TCP New Reno have been studied in [17]. This paper represents throughput results from an experimental study of TCP source variants, Tahoe, Reno and New Reno. In this paper, throughput of each variant is measured by considering the network parameters such as, bandwidth, packet size and congestion window size. The influence of number of burstifiers on TCP performance for an OBS network has been investigated in [18]. This paper shows by increasing number of burst assemblers, TCP goodput grows Effects of several flow-aware and flow-unaware mechanisms have been studied in [19]. This paper shows using these mechanisms improve TCP fairness over OBS networks. In [20] a mechanism for dealing with multicast routing overhead for TCP over Optical Burst Switching networks has been studied. This scheme is based on specialized nodes in All-Optical Networks (AONs) called as Virtual Source (VS) nodes. The goal of this mechanism is increasing the capabilities of virtual source nodes in order to reduce number of failed requests, thus reducing number of bursts retransmission required. The interworking between different access network and an Optical Burst Switched network is considered in [21] that presents the influence of assembly timeout in different access contexts. There are some other modifications in [22-30].

In this paper we propose a fuzzy-based scheme to distinguish cause of the burst drop in the OBS network. We consider two fuzzy sets i.e. congestion drops set and contention drops set. When a burst drop happens in the network, we measure RTTs of the flow packets and then compare them with the minimum RTT (BaseRTT) seen by the flow. We propose a fuzzy membership function that employs RTT and BaseRTT of the flow to represent the degree of **belonging** of the burst drop to congestion drops set and contention drops set. Extensive simulation studies show that the proposed method outperforms other TCPs such as TCP Vegas, TCP Sack and TCP Reno, in terms of throughput, Packet Delivery Count (PDC) and fairness.

The rest of paper is organized as follows. Section 2 describes TCP Vegas as background for the research. Section 3 presents the fuzzy logic and the proposed scheme. Packet level simulation results come in section 4 and finally concluding remarks are given in section 5.

II. TCP VEGAS

The proposed scheme could be implemented in OBS networks that are running under TCP Vegas for improving throughput in these networks. In this section we are going to explain background information about TCP Vegas.

A. TCP Vegas

The proposed scheme is for OBS networks that running under TCP Vegas protocol, so we need to investigate Vegas as a background to the new scheme. For estimating available bandwidth and congestion status in the network TCP Vegas [7,31,32] measures the RTT of each packet in the network. For better understanding of Vegas, we compare Vegas and Reno. TCP Vegas and TCP Reno are different in the slow start, congestion avoidance and retransmission phase.

The first phase that we are going to compare is congestion avoidance. In TCP Reno each packet loss indicates congestion in the network. It means before a packet loss, TCP Reno cannot detect congestion in the network and for congestion detection, TCP Reno has to wait for a packet to drop. On the other hand, in TCP Vegas we have estimated and measured throughput in a specific time window. For detecting congestion in the network, TCP Vegas compares these two throughputs. For determining expected throughput, TCP Vegas measures minimum RTT, which is primarily equaled propagation delay and the queuing delay, then expected throughput could be calculated by the following equation:

$$Expected = \frac{cwnd}{BaseRTT} \tag{1}$$

In this equation cwnd is the current congestion window size. After calculating expected throughput, it is time for actual throughput because we need both expected and actual throughput in TCP Vegas for detecting congestion in the network. For calculating actual throughput, TCP Vegas needs recent RTTS. Actual throughput is calculated by using the following equation:

$$Actual = \frac{cwnd}{RTT} \tag{2}$$

Then Vegas compares expected and actual throughput and calculates difference between these two quantities (denoted as Diff):

$$\text{Diff} = \text{Expected} – \text{Actual}= (\frac{cwnd}{BaseRTT} - \frac{cwnd}{RTT}) \text{ BaseRTT}= \text{cwnd } (1 - \frac{BaseRTT}{RTT}). \tag{3}$$

This non-negative value (Diff) is used to adjust next cwnd size. In Vegas we have two threshold values, denoted as α and β for changing next size of cwnd. TCP Vegas uses these two thresholds and next congestion window is set as follow:

$$cwnd = \begin{cases} \text{cwnd } + 1 & \text{diff } < \alpha \\ \text{cwnd} & \alpha \le \text{ diff } \le \beta \\ \text{cwnd} - 1 & \text{diff } > \beta. \end{cases} \tag{4}$$

When actual throughput is much smaller than expected, it means network may be congested and we must decrease flow rate. On the other hand, when the actual throughput is close to expected, it means available bandwidth in the network is not used efficiently and we must increase flow rate [15].

This behavior is source of a problem in the optical switching network. Sometimes in the network actual and expected throughput are close, and a burst contention happens in the network, in this case Vegas assumes this contention as congestion and starts decreasing its flow rate instead of its increasing.

After congestion avoidance phase we want to investigate slow-start phase in both protocols. TCP Vegas and TCP Reno are different in slow-start phase. In TCP Reno cwnd is increased by one when an acknowledgement is received successfully for a packet. In TCP Reno cwnd is doubled in each RTT. This procedure continues until a packet loss occurs, or exceeding receiver's set cwnd or initial threshold. This method has some downsides. One of them is setting the value of initial threshold. If it is set small, TCP Reno stops increasing cwnd early and it causes low throughput. If the threshold is set too large, TCP Reno continues increasing cwnd for a long time and it leads to exhausting available bandwidth and more packet loss. In the slow-start, TCP Vegas increases cwnd every other RTT, when the cwnd reaches slow-start threshold, Vegas exits slow-start. For making valid comparison between actual and expected throughput in the network, cwnd stays fixed during two consecutive RTT [15].

The last phase to compare between Vegas and Reno is packet retransmission. In this phase like two other phases TCP Vegas and TCP Reno are completely different. TCP Reno enters to retransmission phase, when a timeout happens. It means before happening a timeout TCP Reno continues in loss free phase and after happening a timeout, TCP Reno exits loss free phase and enters retransmission phase. The procedure for TCP Vegas is different. In two situations TCP Vegas enters retransmission phase. When a sender in TCP Vegas receives acknowledgement (ACK), calculates and records the estimated RTTs using current time and the timestamps recorded for the associated packet. Now when a duplicated ACK is received, if the current time is greater than the associated timestamp, for that packet TCP Vegas retransmits the packet. After retransmission, when an ACK is received, and it is first or second one after retransmission, Vegas checks passed time and if it is larger than timeout, retransmits the packet [14].

Fig.1 shows the diagram for TCP Vegas. In Fig.1 A-B is slow-start phase, B-C is transition phase. In the transition phase TCP Vegas goes to loss-free phase from slow-start phase. C-D is loss-free phase and if a packet loss happens, TCP Vegas enters the last phase.

III. Fuzzy Vegas Over OBS

This section explains fuzzy logic and then brings the proposed fuzzy-based congestion control in the optical burst switching networks.

A. Fuzzy Logic

Fuzzy logic is a superset of Boolean logic that it was first introduced by L.Zadeh in the 1960s to handle the concept of partial truth. By using Fuzzy logic we can model the uncertain systems. One of the most important elements of the Fuzzy system is fuzzifier (toward fuzzy sets). The duty of the fuzzifier is mapping discrete (also called crisp) input data into proper values in the fuzzy logic system. This mapping is done by membership function (fuzzy sets). Fuzzy sets provide a transition from false to true (0 to 1). Mathematically, a membership function associates each element of $\mu X(x)$ in the universe of discource U with a number in the interval [0,1] as shown in equation (5):

$$\mu X : \mathrm{U} \rightarrow [0, 1] \qquad (5)$$

So, a fuzzifier maps discrete input data $X \in U$, into a fuzzy set $X \in U$, and $\mu X(x)$ gives the degree of membership of x to the fuzzy set X [35].

Fuzzy logic is based on the theory of fuzzy sets, where an object's membership of a set is gradual rather than just member or not a member. Fuzzy logic uses the whole interval of real numbers between zero (*False*) and one (*True*) to develop logic as a basis for rules of inference. Particularly the fuzzyfield version of the variables enables computers to make decisions using fuzzy reasoning rather than exact [33].

Fuzzy sets are indeed an extension of the classical sets in which only full membership or no membership exist. Fuzzy sets, allows partial membership which 1 denotes full membership and 0 denotes no membership, The numbers from 0 to 1 specifies the membership degree [35].

In fuzzy logic we need to establish a membership function (membership degree). One of the most important challenges is to find a proper membership function by using the fuzzy set theory. Therefore the mathematical model of fuzzy should be established. The numbers from 0 to 1 specifies the membership degree [33].

In our scheme membership function is established by using RTT and BaseRTT, as will be explained later. Here number 1 represents that the network is not heavily congested and 0 shows there is heavy traffic in the network.

After Fuzzy sets, we need fuzzy rules and fuzzy reasoning. It means after establishing membership function we need some rules to operate our system. The amount of the rules depends on the system that we are working on it and these rules specify that the obtained numbers from fuzzy membership go to which one of the domains [35]. In our system we need 3 domains (congested, Normal and empty) so, we need 3 rules, as will be explained later.

B. Fuzzy Vegas over OBS Networks

A network system is a large complex system and there are a lot of parameters in it. We cannot predict behavior of this large system and it has time varying and chaotic one. Besides, TCP face with an important problem when runs over the OBS networks, since it was not designed in OBS networks, rather TCP was designed to work in wired networks where a packet loss can safely be associated to network congestion, so TCP assumes congestion is the cause of any loss in the network and responds to it by slowing down its sending rate [15]. In OBS network we have jointly the congestion and the contention problems which both of them can lead to the packet drop event. Contention drops may lead TCP to long period of inactivity due to its backoff mechanism and it causes the TCP end-to-end throughput will be impaired. We can solve this problem by finding the real cause of every loss. By existing congestion and contention in the network, belonging each packet loss to one of them and the need to separate these two situations it means there is an inherent fuzziness and a partial truth in the core of the network. Hence, when a packet is dropped in the OBS network, we need to distinguish cause of the drop to prevent network from the false congestion detection. When the traditional congestion control system has some shortages and cannot work properly in OBS networks, we need to make some changes in the structure of it; for making new congestion control system we need a framework that can work in this indefinite circumstance, because we do not know packet loss belongs to the congestion or contention and we cannot model the system completely. Fuzzy logic that is commonly known as Computational Intelligence can be viewed as alternative way of designing a new congestion control system when it is convenient and effective to build a control algorithm considering OBS networks function. Fuzzy logic can be viewed as a solution of designing feedback controllers in circumstances where rigorous control theoretic approaches cannot be used due to difficulties in obtaining formal analytical model while at the same time some intuitive understanding of process is available. Besides these problems, in the former methods networks rely on messages notifications which they receive from the intermediate nodes to detect congestion or link failures. The main downside of this problem is depending on the lower layer of the protocols to carry the messages. It needs changing the intermediate nodes of the networks. But when we use and end-to-end approaches, there is no need to change the intermediate nodes and cooperating of them. By using Fuzzy method we can deploy an end-to-end approach that has less overhead comparing other schemes. The key idea of our scheme is to monitor the network and record useful data and infer the current state of the network. Round Trip Time (RTT) and BaseRTT are used as indicators of internal state of the network. The rationale here is that BaseRTT is always smaller than RTT, so we can use them to build a membership function.

This section aims to develop a fuzzy logic-based solution for this problem. The first step to develop a fuzzy-based solution is to define fuzzy sets and its membership function. Here, we consider two fuzzy sets i.e congestion drops set and contention drops set. Hence, we need a membership function that can draw a border between these two sets and can give the degree of belonging of a burst drop to congestion drops set or contention drops set. To outline such membership function we consider this fact that when a network faces with the congestion, their flows experience an increasing RTT. Obviously, larger RTT refers to more intense congestion, hence we can use ratio of BaseRTT to RTT as a congestion measure in the network. As shown in equation (6) this ratio is called congestion indicator (CI) and is used as a fuzzy membership function in the current problem.

$$CI = \frac{BaseRTT}{RTT} \tag{6}$$

Since BaseRTT is always smaller than RTT then we have $0 < CI \leq 1$. When *CI* approaches to *0* it means the network is fully congested and in the opposite side a *CI* approaching to 1 is representing a network with no congestion delay. Other values of *CI* between 0 and 1 stand for a fuzzy network status which is congested only partially.

Now, we employ fuzzy logic to give an improved version of Vegas, called Fuzzy Vegas, for the OBS network. Remember the main goal of this research is differentiating the congestion drops from the contention drops in the OBS networks. Clearly *CI* is the index that can help us in this path. We define two thresholds (minimum threshold or min_{th} and maximum threshold or max_{th}) over *CI* to define three states for the network: congested, normal and empty. Equation (7) shows these three possible situations of the network:

$$\begin{cases} 0 \leq \text{ CI } < \min_{\text{th}} & \mathit{Congested} \\ \min_{\text{th}} \leq \text{ CI } \leq \max_{\text{th}} & \text{Normal} \\ \max_{\text{th}} < CI \leq 1. & \mathit{Empty} \end{cases} \tag{7}$$

As we know, TCP Vegas records always RTT and BaseRTT and hence can compute *CI* easily. When a burst drop occurs, TCP Vegas sees its *CI* and uses Fig. 2 procedure to estimate the network status. When $max_{th}<CI\leq1$, it means RTT is close to BaseRTT and the

burst drop in this situation is more likely caused by the burst contention. When $0 \leq CI < min_{th}$ it means the burst drop in this situation is more likely caused by congestion and hence the congestion window size must be decreased. When $min_{th} \leq CI \leq max_{th}$ happening burst drop can be caused by either congestion or contention, but network is not congested, so we continue to keep *cwnd* fixed. Hence, the congestion windows size can be updated according to Fig. 2 in the OBS network.

It is clear that this algorithm is high-sensitive to its thresholds values. In next section we examine different values for these thresholds to find the best ones.

IV. Packet Level Simulation Results

In order to study performance of the proposed model, we implement Fuzzy Vegas by making some modifications over TCP Vegas module of ns-2 software package and incorporate modules required to implement OBS. In this simulation, we use the network topology shown in Fig. 3 in which there are 16 edge nodes and 3 core nodes. In this network each edge node is connected to the core nodes with a 1ms propagation delay. These links use 100 wavelength channels for transferring data and data rate of each channel is 1Mbps it means each edge node is connected to the core node with a 100Mbps optical link and each link consists 8 wavelength channels for transferring control packets. The link between core nodes consists 100 wavelength channels and data rate of each channel is 100Mbps it means core nodes are connected to each other with 10Gbps bandwidth and the links have 8 wavelength channels for transferring control packets. Data rate of each wavelength channel is 1Mbps. We use mixed time/length based burst assembly algorithm. The burst timeout is 5 ms and the maximum burst size is 50 Kb. The control header processing time is set to 1μs and the offset time is 20 ms. The contention probability varies in the range of [10^{-5}, 10^{-2}] and the simulation time is set to 1000.

A. Impact of min_{th} and max_{th} on Fuzzy Vegas Performance

Obviously, the values of min_{th} and max_{th} have considerable effects on the network performance. In this section, we examine the network performance under different values of min_{th} and max_{th} to find those thresholds that exhibit the best performance. To this end we simulated the network with more than thirty different minimum and maximum thresholds. To keep in summary, we bring here only the simulation results of the following cases:

Case 1: min_{th} =0.3 and max_{th} =0.5
Case 2: min_{th} =0.3 and max_{th} =0.6
Case 3: min_{th} =0.3 and max_{th} =0.7
Case 4: min_{th} =0.3 and max_{th} =0.8
Case 5: min_{th} =0.3 and max_{th} =0.9

Fig. 4 shows the throughputs for each thresholds set. This figure shows that the values of thresholds have a great effect on the performance of Fuzzy Vegas. As can be found in this figure, the best results are achieved for min_{th}=0.3 and max_{th}=0.5. Hence, we accept these values for Fuzzy Vegas' thresholds in following simulations.

B. Fuzzy Vegas Performance

In this section we study Fuzzy Vegas performance in comparison with other TCP implementations. For this purpose we consider again the OBS network of Fig. 3 and simulate it once under Fuzzy Vegas (min_{th}=0.3 and max_{th}=0.5) and then under TCP Vegas, TCP Reno and TCP Sack to compare their performance in terms of throughput, PDC, Fairness and Packet Loss Ratio (PLR). Figure 5 shows throughput of various algorithms. According to this figure, Fuzzy Vegas is much better than other TCP flavors. For example, when the contention probability is 10^{-5}, Fuzzy Vegas' throughput is about 100% higher than conventional Vegas, 180% higher than TCP Reno and 700% higher than TCP Sack.

Packet Delivery Count (PDC) is the other important parameter that has to be analyzed in the OBS network. Figure 6 shows PDC of Fuzzy Vegas and other TCP implementations. This figure shows that Fuzzy Vegas can deliver more packets in compare with other algorithms. Since Fuzzy Vegas can estimate more accurately the network status, hence its congestion control mechanism leads to improved performance in compare to other TCP implementations.

As we know, fairness is an important feature for evaluating of any congestion control schemes. To measure fairness, we use Jain fairness index which is defined as in equation (8):

$$(\sum_{i=1}^{n} \mathbf{B_i})^2 / \mathbf{n} \sum_{i=1}^{n} \mathbf{B_i}^2 \quad (8)$$

Where n is the number of competing flows and B_i is the throughput of the ith flow.

Fig. 7 shows fairness of Fuzzy Vegas and other TCP implementations. Results show that fairness of Fuzzy Vegas is better than other TCP implementations.

And finally, Packet Loss Ratio (PLR) is another important feature which is given in Fig. 8.

According to this figure we can find out that the lowest PLR is for conventional Vegas and next position is for Sack, Reno and then Fuzzy Vegas respectively.

When destination nodes in the Fuzzy Vegas receive more than twice packets compared by conventional Vegas, there is a big chance that PLR for the Fuzzy Vegas is really worse than Vegas, but Fig. 8 shows that the difference is a small number and considering the Fuzzy Vegas throughput, this PLR is acceptable.

V. Conclusion

False congestion detection in the OBS networks is a conventional problem. In this paper we proposed a Fuzzy-based solution that can differentiate the congestion drops from contention drops. For this purpose we employed a membership function that uses both RTT and BaseRTT to estimate the network congestion level. Simulation results in ns-2 environment showed that using fuzzy logic to control congestion in OBS network improves performance of the network.

References

[1] A.Jain, S. Floyd, M.Allman, P.Sarolan Quick-start for TCP and IP, ICSI, 2006.

[2] C. Jin, D. Wei, S. Low, FAST TCP: motivation, architecture, algorithms, performance, INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies (Volume:4) , March 2004.

[3] S. Hegde, et al., FAST TCP in high-speed networks: an experimental study, in: Proceedings, GridNets, Engineering & Applied Science, Caltech, the First International Workshop on Networks for Grid Applications, 2004.

[4] L. Xu, K. Harfoush, I. Rhee, Binary increase congestion control (BIC) for fast long-distance networks, in: Proceedings, INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies (Volume: 4), March 2004.

[5] W. Stevens, TCP slow start, congestion avoidance, fast retransmit,and fast recovery algorithms, RFC, 1997.

[6] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, TCP selective acknowledgement options, RFC, 1996.

[7] L. Brakmo, L. Peterson, TCP Vegas: end-to-end congestion avoidanceon a global internet, IEEE Journal on Selected Areas in Communication, 1995.

[8] D. Katabi, M. Handley, C. Rohrs, Congestion control for high bandwidth-delay product networks, ACM SIGCOMM Computer Communication, PA, 2002.

[9] Óscar González de Dios, Ignacio de Miguel , Ramón J. Durán, Juan Carlos Aguado2, Noemí Merayo2, PatriciaFernández, Impact of TCP Synchronization on Capacity Dimensioning of Optical Burst Switched (OBS) Links, Networks and Optical Communications (NOC), 2012.

[10] Y.W. Wang, Using TCP congestion control to improve the performance of Optical Switched Networks, Communications, 2003. ICC '03. IEEE International Conference on (Volume:2), 2002.

[11] X. Yu, C. Qiao, Y. Liu, TCP implementations and false time out detection in OBS networks, Infocom, 2004.

[12] Q. Zhang, V. Vokkarane, Y. Wang, J.P. Jue, Analysis of TCP over optical burst-switched networks with burst retransmission, in: Proceedings, IEEE GLOBECOM, St. Louis, MO, November 2005.

[13] Q. Zhang, V. Vokkarane, Y. Wang, J.P. Jue, Evaluation of burst retransmission in optical burst-switched networks, in: Proceedings, 2nd International Conference on Broadband Networks, BROADNETS, Boston, MA, 2005.

[14] C. Hsu, T. Liu, N. Huang, Performance analysis of deflection routingin optical burst-switched networks, INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE (Volume:1),New York, NY, June 2002.

[15] BasimShihada, Qiong Zhang, Pin-Han Ho, Jason P. Jue, A novel implementation of TCP Vegas for Optical Burst Switched networks, Optical Switching and Networking, 2010.

[16] Julie Sullivan, Neal Charbonneau, and Vinod M. Vokkarane, Performance Evaluation of TCP over Optical Burst Switched (OBS) Networks using Coordinated Burst Cloning and Forward Segment Redundancy, IEEE Globecom 2010 proceedings, 2010.

[17] Sodhatar, S.H., Patel, R.B. , Throughput Based Comparison of Different Variants of TCP in Optical Burst Switching (OBS) Network , Communication Systems and Network Technologies (CSNT), 2012 International Conference on 2012.

[18] Gurel, G. ; Bilkent Univ., Ankara ;Karasan, E. , Effect of Number of Burst Assemblers on TCP Performance in Optical Burst Switching Networks, Broadband Communications, Networks and Systems, 2006. BROADNETS 2006. 3rd International Conference on, 2006.

[19] Sullivan, J. , Ramos, P. ,Vokkarane, V.M. , Unfairness in TCP performance over lossy optical burst-switched (OBS) networks, Advanced Networks and Telecommunication Systems (ANTS), 2009 IEEE 3rd International Symposium on, 2009.

[20] Sreenath, N. , Fernandez,, Terrance Frederick ; Ramachandiran, S. , Performance analysis of VS nodes in TCP over optical burst switched multicast networks, Emerging Trends in Science, Engineering and Technology (INCOSET), 2012 International Conference on , 2012.

[21] Casoni, M., Raffaelli, C. , TCP Performance Over Optical Burst-Switched Networks With Different Access Technologies, Optical Communications and Networking, IEEE/OSA Journal of (Volume:1 , Issue: 1) , 2009.

[22] Shihada, B. ,Pin-Han Ho ,Fen Hou,Xiaohong Jiang, et al., BAIMD: A Responsive Rate Control for TCP over Optical Burst Switched (OBS) Networks, Communications, 2006. ICC '06. IEEE International Conference on (Volume:6), 2006.

[23] ShupingPeng , Zhengbin Li , Xinlei Wu , AnshiXu, TCP Window Based Dynamic Assembly Period in Optical Burst Switching Network, Communications, 2007. ICC '07. IEEE International Conference on, 2007.

[24] Raffaelli, C., Zaffoni, P. , Simple Analytical Formulation of the TCP Send Rate in Optical Burst-Switched Networks, Computers and Communications, 2006. ISCC '06. Proceedings. 11th IEEE Symposium on , 2006.

[25] Jayaraj, A. , Venkatesh, T. ; Murthy, C.S.R, Loss classification in optical burst switching networks using machine learning techniques: improving the performance of TCP, 2008.

[26] Shihada, B. , Qiong Zhang ; Pin-Han Ho, Threshold-based TCP Vegas over Optical Burst Switched Networks, Computer Communications and Networks, 2006. ICCCN 2006. Proceedings.15th International Conference on, 2006.

[27] Pleich, R. ; Siemens AG, Munich, Germany ,de Vega Rodrigo, M. , Gotz, J. , Performance of TCP over optical burst switching networks, Optical Communication, 2005. ECOC 2005. 31st European Conference on (Volume:4) 2005.

[28]Qiong Zhang , Vokkarane, V.M. ; Wang, Yuke ; Jue, J.P. ,Analysis of TCP over optical burst-switched networks with burst retransmission, Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE (Volume:4), 2005.

[29] Bimal, V. , Venkatesh, T. ; Murthy, C.S.R. , A Markov Chain Model for TCP NewReno Over Optical Burst Switching Networks, Global Telecommunications Conference, 2007. GLOBECOM '07. IEEE, 2007.

[30] Zhu, L. , Ansari, N. ; Liu, J. , Throughput of high-speed TCP in optical burst switching networks, Communications, IEE Proceedings- (Volume:152 , Issue: 3), 2005.

[31] J. Mo, R. La, V. Anantharam, J. Walrand, Analysis and comparison of TCP Reno and Vegas, , INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE (Volume:3) , March 1999.

[32] E. Weigle, W. Feng, A case for TCP Vegas in high-performance com-putational grids, in: Proceedings, 10th IEEE International Sympo-sium High Performance Distributed Computing, San Francisco, CA, August 2001.

[33] JunhaiLuo, Xue Liu, Mingyu Fan, A trust model based on fuzzy recommendation for mobile ad-hoce networks, Computer Networks, Local Computer Networks, 2008. LCN 2008. 33rd IEEE Conference on 2009.

[34] DE OLIVEIRA, R, BRAUN, T., A DELAY-BASED APPROACH USING FUZZY LOGIC TO IMPROVE TCP ERROR DETECTION IN AD HOC NETWORKS, 2004. WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, 2004. WCNC. 2004 IEEE .

[35] Ramakrishna Shenai, Sunil Gowda and Krishna M Sivalingam, Washington state University, 2001.

[36] Martin Levesque and Halima Elbiaze, Graphical probabilistic routing model for OBS networks with realistic traffic scenario, Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE , 25 Jul 2009.

[37] B. Praveen, J. Praveen, C. Siva Ram Murthy, On using forward error correction for loss recovery in Optical Burst Switched networks, Computer Networks: The International Journal of Computer and Telecommunications Networking, 2007.

[38] Vasco N.G.J. Soares, Iuri D. C. Veiga and Joel J. P. C. Rodrigues, OBS simulation tools: a comparative study, Communications Workshops, 2008. ICC Workshops '08. IEEE International Conference on, 2008.

[39] Sunil Gowda, Ramakrishna K Shenai, Krishna M Sivalingam and HakkiCandanCankaya, Performance Evaluation of TCP over Optical Burst-Switched (OBS) WDM networks, Communications, 2003. ICC '03. IEEE International Conference on (Volume:2), 2003.

[40] GurayGuel, OnurAlparslam and EzhanKarasan, nOBS: an ns2 based simulation tool for performance evaluation of TCP Traffic in OBS networks, Annales Des Télécommunications, 2007.

[41] Ns-2, Network Simulator, www.isi.edu.

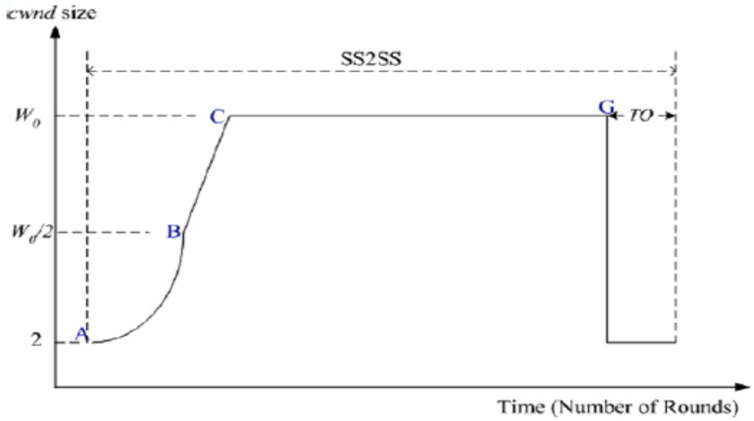


Fig. 1. TCP Vegas different phases [15]

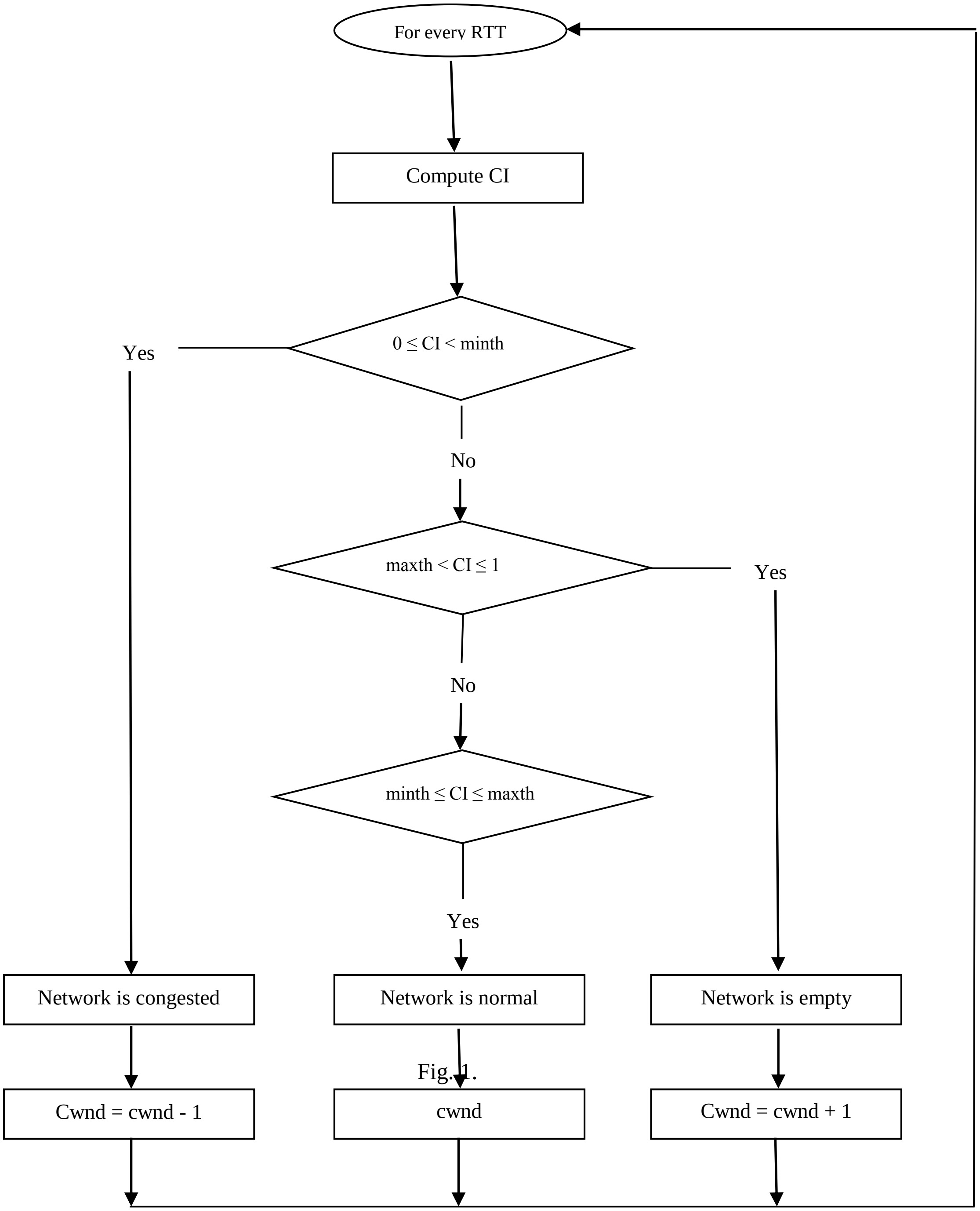


Fig. 2. The flowchart for the fuzzy-based TCP Vegas congestion control algorithm

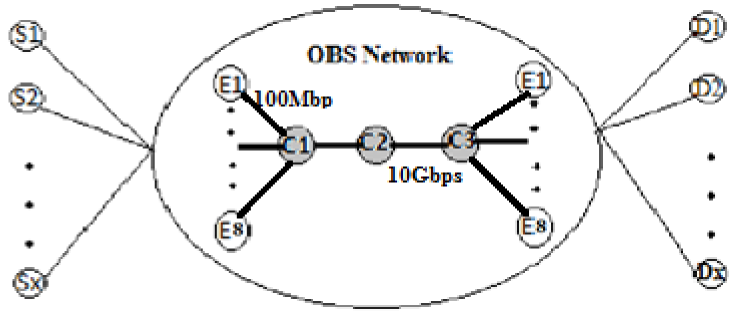


Fig. 3. The network topology adopted in the simulation

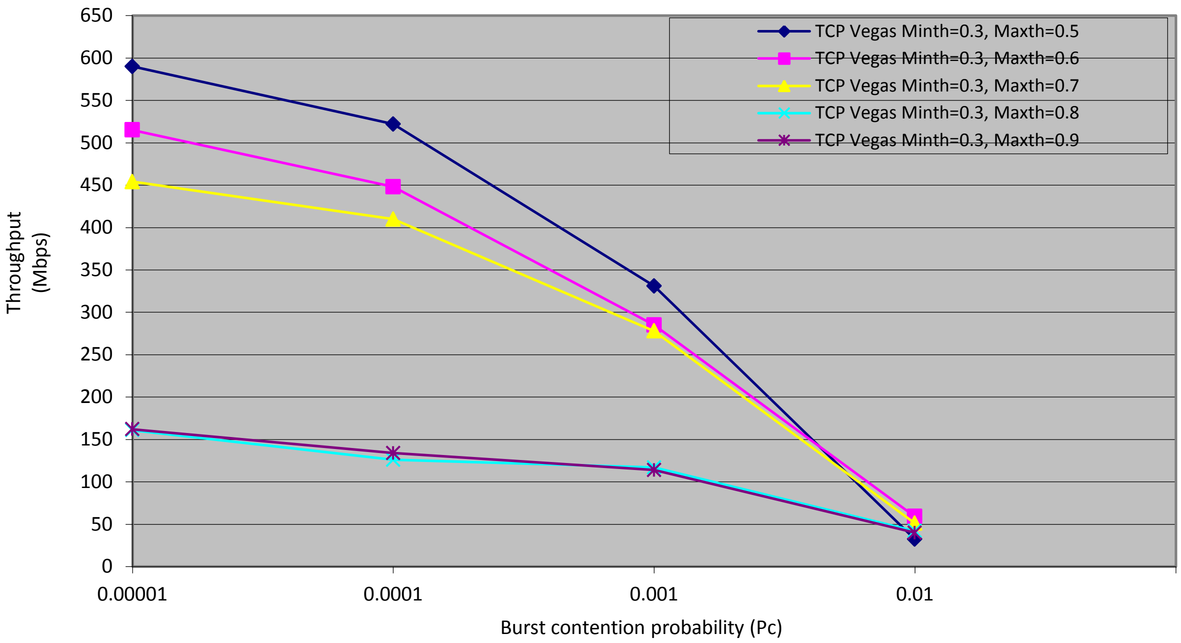
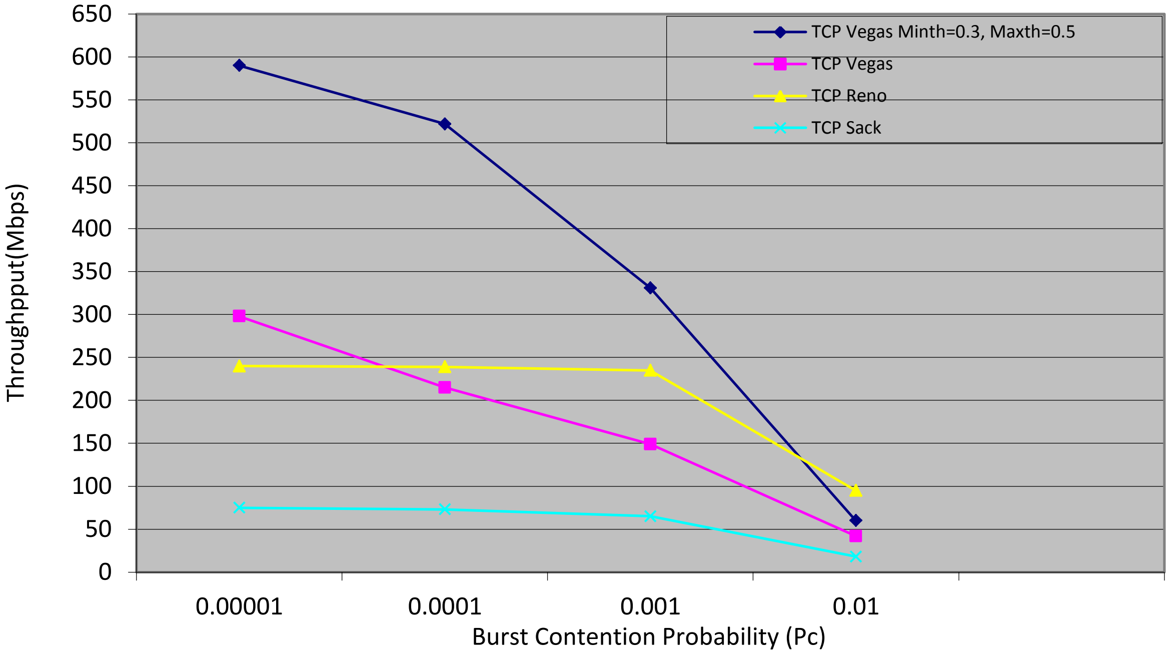


Fig. 4. Throughput comparison of Fuzzy Vegas with different thresholds



.

Fig. 5. Throughput comparison of TCP implementations and Fuzzy Vegas

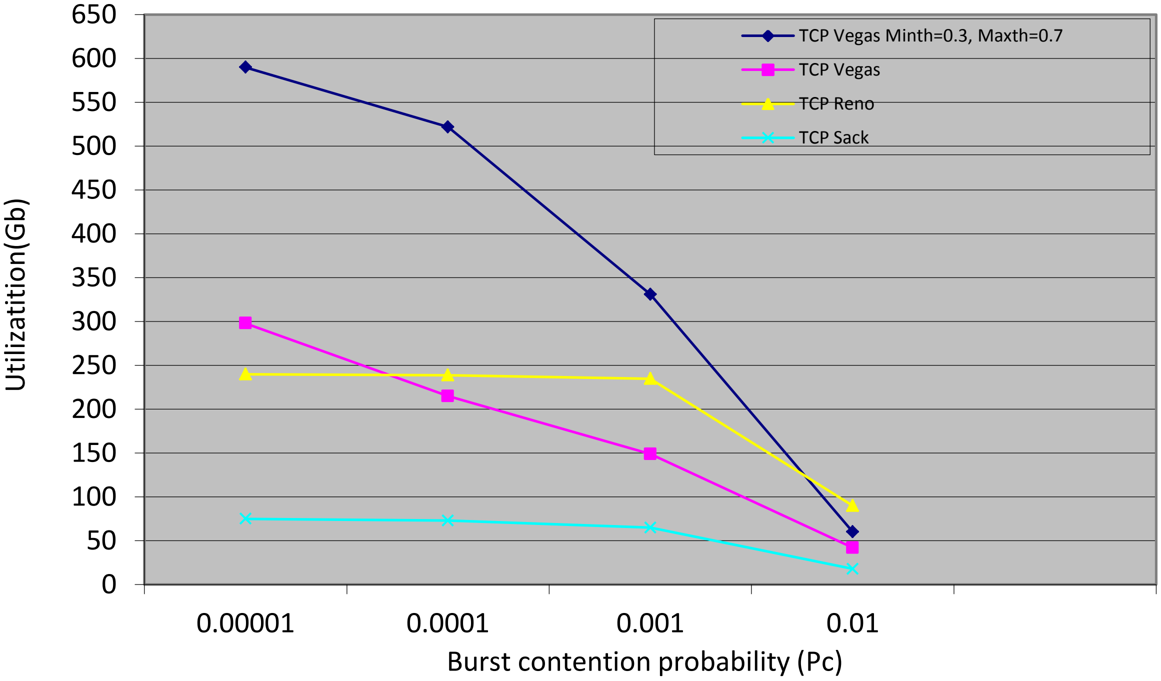


Fig.

6. Comparison of the Packet Delivery Count between TCP implementations and Fuzzy Vegas

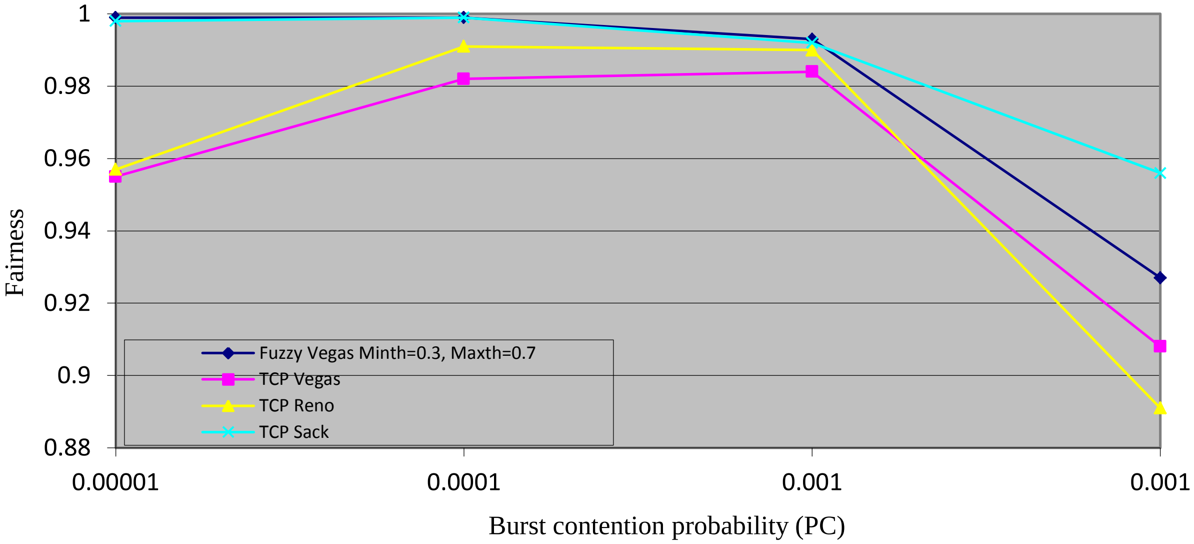


Fig. 7. Comparison of the fairness between TCP implementations and Fuzzy Vegas

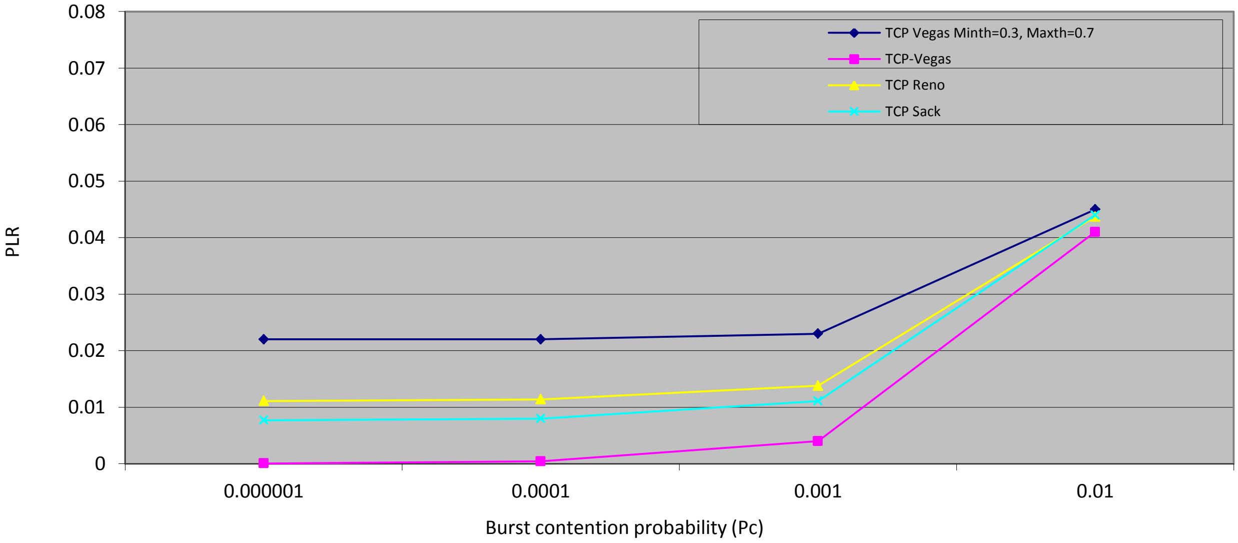


Fig. 8.Comparison of the PLR between TCP implementations and Fuzzy Vegas

Improved Algorithm for fusion of Satellite Images Using Combined DWT-FDCT Transforms

Manjushree B S(Author 1)
Student, CSE, DBIT/VTU
Bangalore, India

Shruthi G(Author 2)
Asst.Prof, CSE, DBIT/VTU
Bangalore, India

Abstract
Image fusion based on the Fourier and wavelet transform methods retain rich multispectral details but less spatial details from source images. Wavelets perform well only at linear features but not at non linear discontinuities because they do not use the geometric properties of structures. Curvelet transforms overcome such difficulties in feature representation. A novel fusion rule via high pass modulation using Local Magnitude Ratio (LMR) in Fast Discrete Curvelet Transforms (FDCT) domain and Discrete wavelet transforms (DWT) is defined. For experimental study of this method Indian Remote Sensing (IRS) Geo satellite images are used for Pan and MS images. This fusion rule generates HR multispectral image at high spatial resolution. This method is quantitatively compared with Wavelet, Principal component analysis (PCA) fusion methods. Proposed method spatially outperforms the other methods and retains rich multispectral details.

Key words:
Image Fusion, Fast Discrete Curvelet Transforms, Discrete wavelet transforms, Local Magnitude Ratio (LMR)

I. INTRODUCTION

Image fusion integrates the multisensor data to create a fused image containing high spatial, spectral and radiometric resolutions. In remote sensing, image fusion is most valuable technique for utilization of multisensor, multispectral at various resolutions of earth observation satellites [6]. Spatial resolution plays a vital role to delineate the objects in the remote sensing image. It is easy to interpret the features with high spatial resolution [12] image with multispectral information than the single high resolution Pan image. Image fusion enhances the spatial, spectral and radiometric [7] resolutions of images. Fast Fourier and wavelet transform based image fusion methods retain better spectral characteristics but represent poor spatial details in fused images.

A. Problem statement

Sparse representation of signals is now possible utilizing many different Greedy approaches including Matching Pursuit, Orthogonal Matching Pursuit. These techniques are used to represent signals with the fewest number of non-zero coefficients. Principal Component Analysis is one of the powerful state-of-the-art image fusion approaches in terms of visual inspection and quantitative evaluation metrics. This fusion is carried out by integrating the principal components of images to be fused. PCA is one of the special domain fusion approach. DWT is one of the transform domain fusion approach. In DWT cost of computing is high and takes long compression time. Both PCA and Sparse fusion have specific advantages and disadvantages. PCA fusion will enhance the spatial quality but have dense nonzero entries that might represent uninformative features. Sparse fusion preserves important information but high spatial resolution is lacking. An algorithm which utilizes the advantages of both PCA and Sparse representation for fusing common and innovative features of the captured images is proposed. This algorithm also overcomes the disadvantages of both PCA and Sparse representation. The effectiveness of proposed method by comparing its results with PCA and Sparse Fusion is demonstrated.

II. RELATED WORK

Image Fusion is used extensively in image processing systems. Various Image Fusion methods have been proposed in the literature to reduce blurring effects. Image fusion enhances the quality of image by removing the noise and the blurriness of the image. Image fusion takes place at three different levels i.e. pixel, feature and decision. Its methods can be broadly classified into two that is special domain fusion and transform domain fusion. Averaging, Brovery method, Principal Component Analysis (PCA), based methods are special domain methods. But special domain methods produce special distortion in the fused image .This problem can be solved by transform domain approach. The multi-resolution analysis has become a very useful tool for analyzing images.

A brief summary of the literature is given below:

H. Yin, S. Li, (2011) [3] proposes a novel multimodal image fusion scheme based on the joint sparsity model which is derived from the distributed compressed sensing. First, the source images are jointly sparsely represented as common and innovation components using an over-complete dictionary. Second, the common and innovations sparse coefficients are combined as the jointly sparse coefficients of the fused image. Finally, the fused result is reconstructed from the obtained sparse coefficients. Furthermore, the proposed method is compared with some popular image fusion methods, such as multi scale transform-based methods and simultaneous orthogonal matching pursuit-based method. The experimental results demonstrate the effectiveness of the proposed method in terms of visual effect and quantitative fusion evaluation indexes.

O, R et al. (1997) [4] has discussed a novel approach for the fusion of spatially registered images and image sequences. The fusion method incorporates a shift invariant extension of the discrete wavelet transform, which yields an over complete signal representation. The advantage of the proposed method is the improved temporal stability and consistency of the fused sequence compared to other existing fusion methods. Information theoretic quality measure based on mutual information to quantify the stability and consistency of the fused image sequence is introduced.

Li, H et al. (1995) [5] has discussed that here, the wavelet transforms of the input images are appropriately combined, and the new image is obtained by taking the inverse wavelet transform of the fused wavelet coefficients. An area-based maximum selection rule and a consistency verification step are used for feature selection. A performance measure using specially generated test images is also suggested.

III. METHODOLOGY

Objective of this paper is to develop a method, which retains better characteristics of both spatial and spectral qualities of source images. Wavelet transforms do not represent the curved objects as in HR Pan image. Curvelet transforms overcome such difficulties of wavelet. Over a period, curvelet transforms are evolved in two generations, such as first generation curvelet transforms and second generation curvelet transforms named as Fast Discrete Curvelet transforms (FDCT). First generation Curvelet transforms computational complexity is more to compute the curvelet coefficients [2]. To overcome these difficulties Emmanuel J.Candes [1] developed FDCT. FDCT represents linear-edges and curves accurately than any other mathematical transforms.

DWT is any wavelet transform for which wavelets are discretely sampled. Key advantage of this is it captures both frequency and location information at a time. Here we are using 'Haar' wavelets. This is used to pair up input values, storing the difference and passing the sum. This process is repeated recursively, pairing up the sums to provide the next scale, which leads to 2n-1 differences and a final sum[14].

Lastly the output images of both FDCT and DWT are combined to get a final output image which has better spectral and spatial resolutions than other transforms to which compared here.

A. FDCT

The curvelet coefficients are obtained by using the Plancherel's theorem for $j \geq 0$,

$C_{j,k,l}$=(f, $\varphi_{j,k,l}$)= $\int_{R^2} f(x)\overline{\varphi_{j,k,l}(x)}$ dx= $\int_{R^2} \hat{f}(\omega)\overline{\hat{\varphi}_{j,k,l}(\omega)}d\varepsilon$

Where $C_{j,k,l}$ is the curvelet coefficients at scale *j* and in the direction *l* at location *k.*

The low frequency (coarse scale) coefficients are shown at the center of the display in Fig. 1. The cartesian concentric corona show the coefficients at different scales. The outer corona corresponds to high frequencies. There are four strips associated to each corona in north, south, east and west direction, these are further subdivided in angular panels. Each panel represents coefficients at a specified scale and orientation suggested by the position of the panel.

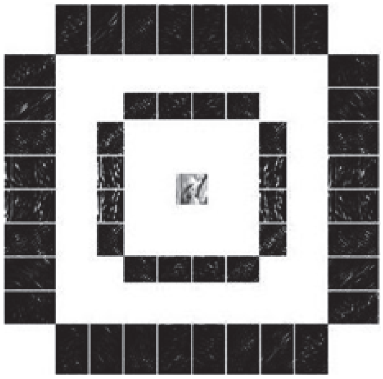


Fig.1 Curvelet coefficents at scale *j*=0 and at scale *j*=1, 2 in multiple directions

B. LMR

In a directional sub-band, bigger curvelet coefficients of HR Pan image and LR multispectral image represent sharp local feature [11]. In this paper, we define a Local Magnitude Ratio (LMR) to inject high frequency details of the local image feature into the fused image. LMR is defined as follows.Let us suppose that *cj,l*(*M*) , *cj,l*(*P*) are the sub-band curvelet coefficients at scale *j* in a direction *l* of the multispectral band *M* and panchromatic image *P* at higher frequencies respectively.

$$LMR_{j,l}(x, y) = \frac{|c_{j,l}(M(x,y))|}{|c_{j,l}(P(x,y))|} \qquad (1)$$

Where $LMR_{j,l}$(x, y) is the sub-band curvelet coefficients at scale *j* in direction *l* at location (x, y).

If $LMR_{j,l}$(x, y) ≤ 1 then *cj,l*(*P*(*x, y*)) represents good local feature. If $LMR_{j,l}$(x, y) > 1 then *cj,l*(*M*(*x, y*)) represents good local feature. Fusion rule to inject high spatial details from HR panchromatic image into LR multispectral image bands is defined using LMR of curvelet coefficients in the directional high frequency sub-bands.

C. Image fusion algorithm using FDCT

Spatial resolution ratio between HR Pan image and LR multispectral image is 2. Input images size must be power of 2 for coherent multi resolution decomposition in FDCT domain. To obtain HR multispectral image, high frequency details are injected into each LR multispectral band in FDCT domain. The fusion rule based on the LMR in FDCT domain is defined as follows.

1) LR multispectral image is resampled to the scale of HR Pan image in image co registration. i.e., both the images must be at identical geometry and of same size.
2) The multispectral data in Green, Red and near infrared bands are extracted band wise.
3) Apply fast discrete curvelet transform (FDCT) to multispectral band M and Panchromatic image P. The input images are decomposed into four levels in multiple directions. Number of directions depends on the image size and decomposition levels.

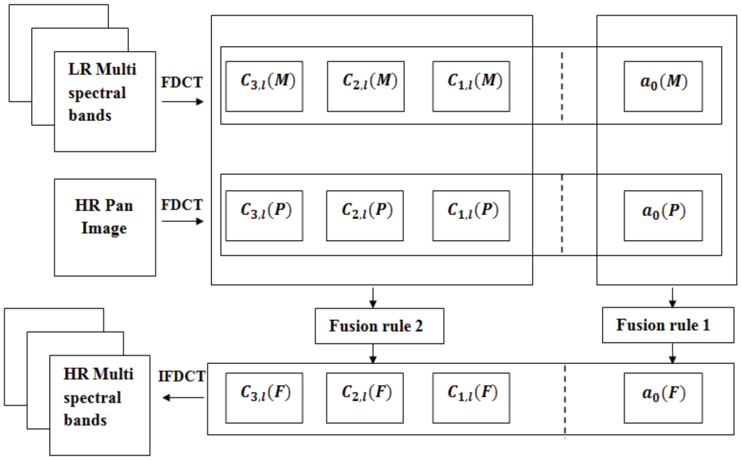
LMS = {*c*3,*l*(*M*), *c*2,*l*(*M*), *c*1,*l*(*M*), *a*0(*M*)}

HPan = {*c*3,*l*(*P*), *c*2,*l*(*P*), *c*1,*l*(*P*), *a*0(*P*)}

where *LMS* is the set of curvelet coefficients for low resolution multispectral band, where *HPan* is the set of curvelet coefficients for high resolution panchromatic image and *a*0(*M*) is the coarser scale coefficients of the multispectral band M, similarly *a*0(*P*) for the panchromatic image *P*.
4) Fusion rule 1 is defined for the curvelet coefficients at lower frequencies (coarser scale coefficients). Construct coarser scale coefficients for fused image *F* from LR multispectral band *M* such that *a*0(*F*) = *a*0(*M*)
5) Fusion rule 2 is defined for the curvelet coefficients at higher frequencies based on high pass modulation. Construct the multidirectional multiresolution curvelet coefficients *cj,l*(*F*) by using Equation for fused image.
6) Construct the set of curvelet planes for fused image as

HFus = {*c*3,*l*(*F*), *c*2,*l*(*F*), *c*1,*l*(*F*), *a*0(*F*)}

and apply the Inverse Fast Discrete Curvelet Transforms (IFDCT).
7) Apply steps (3) to (6) for each multispectral band.
8) Combination of three resultant fused bands provide the HR multispectral fused image.



$$C_{j,l}(F(x, y)) = \begin{cases} C_{j,l}(P(x,y)) * LMR_{j,l}(x,y) & \text{if } LMR_{j,l}(x,y) > 1; \\ C_{j,l}(P(x,y)) & \text{if } LMR_{j,l}(x,y) \le 1; \end{cases} \qquad (2)$$

D.DWT

The basic idea of DWT for one-dimensional signals is briefly described. A signal is split into two parts, usually the high frequency and the low frequency part. This splitting is called decomposition. The edge components of the signal are largely confined to the high frequencies part. The signal is passed through a series of high pass filters to analyze the high frequencies, and it is passed through a series of low pass filters to analyze the low frequencies. Filters of different cutoff frequencies are used to analyze the signal at different resolutions. Let us suppose that x[n] is the original signal, spanning a frequency band of 0 to π rad/s. The original signal x[n] is first passed through a halfband highpass filter g[n] and a lowpass filter h[n]. After the filtering, half of the samples can be eliminated according to the Nyquist's rule, since the signal now has the highest frequency of π/2 radians instead of π. The signal can therefore be subsampled by 2, simply by discarding every second sample. This constitutes one level of decomposition.

The above procedure can be repeated for further decomposition. The outputs of the highpass and lowpass filters are called DWT coefficients and by these DWT coefficients the original image can be reconstructed. The reconstructed process is called the Inverse Discrete Wavelet Transform (IDWT). The above procedure is followed in reverse order for the reconstruction. The signals at every level are upsampled by two, passed through the synthesis filters g'[n], and h'[n] (highpass and lowpass, respectively), and then added. The analysis and synthesis filters are identical to each other, except for a time reversal.Here we use Haar wavelet filters.

An image can be decomposed into a pyramidal structure, which is shown in Figure 2, with various band information: low-low frequency band LL, low-high frequency band LH, high-low frequency band HL, highhigh frequency band HH.[13].

Finally both the output images got after applying FDCT and DWT transforms are combined to get a more quality image output.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Geo satellite images are used for Pan and MS images. For clear visualization, subset images of the fused images of different techniques are shown in Figure 2. Figure 2(a) is the original HR Pan image and Figure 2(b) is the resampled LR multispectral image. Figure 2(c) is the HR multispectral image obtained by fusion rule based on FDCT. Quality of the fused images is evaluated with both spatial and spectral quality measures.

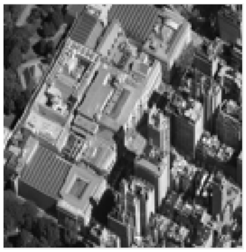


Fig (2a)

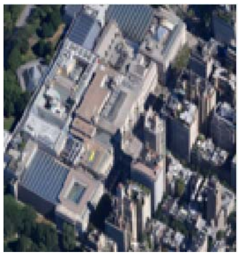


Fig (2b)

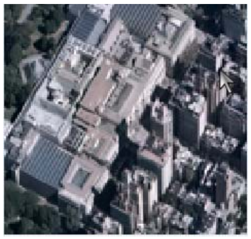


Fig (2c)

A. Spatial Quality Evaluation

Each MS band in a fused image is compared to the HR Pan image for spatial quality evaluation.

1) ***Entropy***:

Entropy is a measure to directly conclude the performance of image fusion. The Entropy can show the average information included in the image and reflect the detail information of the fused image. Commonly, the greater the entropy of the fused image is, the more abundant information included in it, and the greater the quality of the fusion. According to the information theory of Shannon, the entropy of image is defined as E = - $\sum_{i=0}^{n} p_i \log_2 p_i$.Where *E* is the entropy of image and *pi* is the probability of *i* in the image , here *pi* is the frequency of pixel values from 0 to n in the image. We normalized the HR Pan data and LR MS data radiometric resolutions. Entropy values bandwise are shown in Table I.

Table I

Band	FDCT	Wavelet	PCA
1	6.22	5.91	5.66
2	6.78	6.55	6.81
3	6.29	6.15	6.20
Avg	6.43	6.21	6.23

B. Spectral Quality Evaluation

Resampled multispectral bands of sensor image and corresponding bands in the fused image are compared for spectral quality evaluation.

1) ***Spectral Angle Mapper(SAM)***:

Let *v* and ˆ*v* be two vectors having *l* components of resampled multispectral sensor band and the corresponding band in the fused image respectively. Spectral angle mapper (SAM) is the absolute value of the angle between the two vectors [2].

$$\mathrm{SAM}(v,\hat{v}) = \cos^{-1}\left(\frac{\langle v,\hat{v}\rangle}{\|v\|_2 \|\hat{v}\|_2}\right)$$

SAM is measured in either degrees or radians and is usually averaged over the whole image to yield a global measurement of spectral distortion. SAM values equal to zero denotes the absence of spectral distortion. Table II shows the SAM values for each fused band.

Table II

Band	FDCT	Wavelet	PCA
1	0.04	0.07	0.06
2	0.05	0.07	0.21
3	0.03	0.06	0.11
Avg	0.04	0.07	0.13

V. CONCLUSION

We have described new fusion improved method based on FDCT and DWT. Two fusion rules are defined, fusion rule 1 is for curvelet coefficients at lower frequencies and fusion rule 2 in FDCT is for the curvelet coefficients at higher frequencies. Fusion rule 1 substitute the coarser scale coefficients of LR multispectral bands into the coarser scale coefficients of HR Pan image. Fusion rule 2 is based on the high pass modulation using Local Magnitude Ratio (LMR) of the curvelet coefficients in each orientation and scale. Bigger curvelet coefficients of HR Pan image and LR multispectral image represent sharp local feature. LMR directs the injection of high frequency details of the local image feature in HR Pan image into fused image. For experimental study of this method Indian Remote Sensing (IRS) Geo satellite images are used for Pan and MS images. This fusion rule generates HR multispectral image at 2.5m spatial resolution. This method is quantitatively compared with Wavelet, Principal component analysis (PCA) fusion methods. Proposed method spatially outperforms the other methods and retains rich multispectral details.

As of now only FDCT domain is used for fusion. Work is going on for combining both the domains(FDCT-DWT) and obtain more good quality images and to find out more spatial and spectral quality measures.

REFERENCES

[1] Emmanuel.J. Candes, L. Demanet,D.Donoho and L.Ying."FastDiscreteCurvelet transforms(FDCT)",Caltech, Pasadena, CA, March 2006.

[2] F. Nencini, A. Garzelli,S. Baronti,L. Alparone ," Remote sensing image fusion", Informaion Fusion, 8:143-156, May 2006.

[3] Ying. Li, Xing. Xu,Ben-Du. Bai,Y.N. Zhang, " Remote sensing image fusion based on Fast Discrete Curvelet Transform", In International conference on machine learning and cybernetics, Kumming,China, July 2008.

[4] A. Golibagh, Mehran. Yazdi," A novel image fusion method using curvelet transform based on linear dependency test", In International conference on digital image processing, pp:351-354, Bangkok, Thailand, March 2009.

[5] Shutao. Li, Bin. Yang," Multi focus image fusion based on wavelet and curvelet transform", Pattern recognition letters, 29:1295-1301, February 2008.

[6] C.V. Rao, P.S. Reddy, D.S. Jain, K.M.M. Rao," Quantitative value addition analysis of multisensory data fusion" , The ICFAI journal of earth sciences, 1(1):82-96, March 2007.

[7] M.F. Yakhdani, A. Azizi, "Quality assessment of image fusion techniques for multi sensor high resolution satellite images(cas study: IRSP5 and IRSP6 Satellite images)", In ISPRS TC VII Symposium, pp:205-209,Vienna, Austria,July 2010.

[8] Jianwei. Ma, Gerlind. Plonka, "A Review of curvelet and recent applications", Caltech, Pasadena, CA, March 2006.

[9] J. Zhou,D.L. Civico, J.A. Silander,"A Wavelet Transform method to mrege Landsat TM and SPOT Panchromatic data", International Journal of Remote Sensing, 19(4):743-757,1998.

[10] G. Hong, Y. Zhang, "Comparison and improvement of Wavelet-based image fusion", International Journal of Remote Sensing, 29(3):673-691,Dec,2007.

[11] Chengzhi Deng, Hanqiang Cao, Chao Cao, Shengqian Wang,"Multisensor image fusion using fast discrete curvelet transform", Remote Sensing, GIS Data Processing and Applications, Proc. of SPIE Vol. 6790 679004,2007.

[12] C.V. Rao, J. Malleswara Rao, A. Senthil Kumar, and A. S. Manjunath."Restoration of high frequency details while constructing the high resolution image." In India Conference (INDICON), 2011 Annual IEEE, pp. 1-5. IEEE, 2011.

[13] Nataša Terzija, Markus Repges, Kerstin Luck, Walter Geisselhardt, IEEE "DIGITAL IMAGE WATERMARKING USING DISCRETE WAVELET TRANSFORM"

[14] Discrete wavelet transforms-wikipedia the free encyclopedia,"en.m.wikipedia.org".

Biometric Student Record Management System

Onuiri Ernest E
Department of Computer Science, Babcock University
Ilishan-Remo, Ogun State; Nigeria

Yadi Chukwuemeka
Department of Computer Science, Babcock University
Ilishan-Remo, Ogun State; Nigeria

Oludele Awodele
Department of Computer Science, Babcock University
Ilishan-Remo, Ogun State; Nigeria

Oshilagun Ibukun
Department of Computer Science, Babcock University
Ilishan-Remo, Ogun State;

Etuk Otobong
Department of Computer Science, Babcock University
Ilishan-Remo, Ogun State; Nigeria

Abstract – **Information is an important part of any system. In the academic world, information is especially very important and essential. Students have to register for courses, take attendance, quizzes, and exams and as well as check their scores. Years after graduating from the school, students come back asking for transcript. It is therefore very important to handle students' records in a way that is accessible, maintainable and secure. The manual method of cumulating and storing student record is often prone to various degrees of human error and is also unsecured making it exposed to unauthorized personnel. This paper presents the design and development of a biometric student record management that provides an interface between student and the institution to enable prompt checking of grades, as well as track their progress and efficiently record each student's attendance for every lecture attended through the use of a biometric device. The methodology used in developing this system is the waterfall methodology and this was used because it is a one dimensional model, meaning it is very easy to implement and also the documentation is done at the beginning of the software development. During the course of this research, it was realized that developing a biometric student record management system was a herculean task. This system was given to random students to use and 90% of them loved the interactive nature of the system. A projection of record growth in relation to student population and system requirement was carried out in the study.**

Keywords – Fingerprint, Biometrics, Biometric Student Record Management System (BSRMS), Student Information Management System (SIMS), Grade Point Average (GPA), Students

I. INTRODUCTION

Student records are important and form a vital part of the education system of any institution [1]. These institutions are confronted with the burden of monitoring the achievement of individual students as well as show that all students are meeting the set standard for learning. The ability of an educational institution to ease this burden is affected by the institution's access to students' information which has to be accurate and relevant. A well-designed student record yields many benefits. Figures obtained from the National Center for Education Statistics suggests that the most important benefit of a well-designed student record is its ability to report information about individual students, school programs and the institution as a whole to be used for decision making at all levels of management [2]. In the process of grading the students, the records might get lost or destroyed in case of a disaster [3]. Student record system plays a vital role in the overall functioning of the educational system, but more importantly it helps to enhance student growth and meet the needs of students. Student information system deals with all kind of information such as individual student details, student academic reports, institution details, course details and other resource related details too. It keeps track of all the details pertaining to a student from the first day to the last day of a course which can be used for all reporting purposes, tracking of attendance in class, progress in the course and final exam result. A well-designed student record system is one which involves the creation, storage, retrieval, maintenance of students' records to ensure that the records are readily available to an educational institution [4]. However, in many institutions especially in Nigeria, the educational institutions still embrace the manual method of recording and managing students' information which wastes time and is demanding.

A. *Why Fingerprint Biometrics?*

Fingerprints are considered to be the best and fastest method for biometric identification. They are secure to use and unique for every person. It has been proven over the years, through research that no two individuals have the same fingerprint [5]. Fingerprints have been used as a means to identify a person for a long time. Each finger print is made up of pattern of ridges and valley on the surface of a finger tip. Finger ridge configurations do not change throughout the life of an individual except due to accidents such as bruises, cuts on the fingertips or deliberately damaging the fingers [6]. When compared with other biometrics features such as retina, palm or voice, fingerprint-based biometrics is the most proven and commonly used technique which accounts for its large market shares. Fingerprint biometrics are also faster, energy conservative and cheaper than other biometric techniques [7]. It is also affordable to scan and can be used in computers for a lot of applications [8].

Biometric systems have alot of benefits over the traditional systems. Firstly it is impossible to share and very hard to reproduce an individual biometric feature. Biometric systems also eliminates the need to remember and memorize long and random passwords or pins thereby enhancing user convenience. Biometrics systems also provides the same level of security to all users unlike passwords/pins and is repellent to brute force attacks. Moreover, biometrics is one of the few techniques that can be used to determine whether an individual is pretending to be someone else [9].

A biometric student record management system would provide the needed solution. It is a system developed to manage the records of each student using a biometric device as a way of enrolling each student. This system will also help in taking attendance of students and help in evaluating the attendance of students. Our scope emphasizes on three functionalities which are course management, course scheduling and grading.

II. REVIEW OF CLOSELY RELATED WORKS

In this section analysis of some closed related works on student record managemment and biometric attendance management system was carried out. This section discusses the most closely related, then provides a comparative analysis between the related works and our system.

A. *Cinfores Web Portal:*

This software is used in providing services such as information about an institution, various schools and faculties, departments; maintaining accounts; payments through an online portal and payments of records; email services and students exams. It is a web application that helps with student registration, academic course registration, students personal and academic records, academic result approval and CPGA check. [10]

B. *An attendance monitoring system:*

This system is used to track the attendance of employees. It was developed to be accurate, fast and very efficient way of tracking employees. It adopts fingerprint verification by using extraction of minutiae technique. A survey using this system showed that the fingerprint biometric identifier was found suitable for the employee attendance management system of the institution/organization. [11]

C. *Online Biometrics Class Attendance Management System:*

This software is used specifically by Covenant university to mark the attendance of students in class. Fingerprint biometrics is used for processing and managing the class attendance. The system is robust and using trial tests conducted on 60 students in Covenant University, it achieved an average of 89.33% accuracy for first signing attempt. [5]

D. *Biometric-based Attendance System:*

This system is used to take attendance during lecture periods using fingerprint as biometric for LASU Epe Campus. It increases accuracy in attendace-taking, security and also efficiently calculates the attendance percentage. The attendance system makes use of a fingerprint identification system which compares the biometrics of students with every record in its database. [12]

E. *Web-based Student Academic Records Information System:*

This software was used for management and processing of data/information for every student in the school in a seamless and interactive manner. It solves the need of tracking students academic performance progress at each level as well as other managerial activities. The design of the system adopted a client/server technology. The client side was developed with Visual Basic.NET and the server side was designed with MYSQL. This system provides a 3 year performance analysis for students of a given programme. [10]

F. *Biometric Attendance Management System:*

This system uses wireless ZigBee technology. It has an attendance report that will be sent to the respective department HOD or class in-charge once in 15 days. Reports can also be sent to the parents e-mail id. Some also used RFID to track attendance of the students [13][14].

After scrupulous reviewing the aforementioned work, the similarity this system has with ours is:

- It records the students’ data necessary for managerial activities and decision making.
- The system is a single platform that was used for the management and processing of data/information for all students in an interactive and seamless manner. This system increases efficiency in the delivery of service.

Some of the weaknesses that can be highlighted from this work include:

- No means of taking the attendance of students which is a very important student record.
- It doesn’t provide an extra means of security to protect its documents and shield unauthorized personnel from viewing them.

The BSRMS would address these issues by:

- Taking the attendance of students and providing an analysis to view student performance in class.
- Providing an extra means of security by using a biometric device to validate users logging on to the system.
- Providing information that will be used for more than just normal data processing. These processed data will be used in supporting decision making which would be efficient in all managerial activities.
- Allowing generation of the necessary academic documents.

In conclusion, in order to address these issues, the following were done:

- Interviewed different stakeholders including management staffs and students of Babcock University.
- Reviewed other methods and systems used in collecting and managing student information which would help us gain an in-depth understanding of the system and its features and functionalities.
- A one year academic result for 10 students per session was simulated for the department of Computer Science in Babcock University.

III. METHODOLOGY

The waterfall model of software development was adopted to achieve the design of this application within the specified time limits and constraints. It is a sequential design process, often used in the process of developing software in which each progress is viewed as flowing downwards like a waterfall [15]. This is because it explicitly outlines each step and processes associated with developing an application. All system and user requirements are completed before the system design activity proceeds.

System analysis tools such as interviews and on-site observation was adopted in obtaining detailed facts about the current system so as to identify its limitations. Useful information acquired from the current system was applied to meet the necessary requirements and objectives of the proposed system.

The attendance system was implemented using Microsoft Visual Basic and a biometric device. It consists of two stages: the enrolment stage where each student biometric details was taken and stored in the database and the authentication stage where each student's biometric features were extracted and compared with all fingerprint templates in the database.

The integrated web platform of the system was designed into two modules. The first module was designed mainly to manage all processes associated with students registering for courses and lectures recording and storing students' grade scores. The second module was primarily developed to generate student's cumulative assessment result and grade point. The system will also generate the 5% attendance, the Grade Point Average (GPA) the Cumulative Grade Point average (CGPA).

There are numerous benefits to using a Student Records Management. One of the benefits is the use of a central database. This database is the core for all actions in the system and can be easily updated and used to ease all the system's processes. This storage method is more efficient than a paper-based file system. Another factor which the system has taken into consideration is human error made in the recording and filing process which is avertable in a database system. It also makes provision of easy corrections of errors made.

Benefits of the proposed system:

1) *REDUCED TIME CONSUMPTION:*
Reduce the time taken to process the queries of users, get student records for decision making etc.
2) *REDUCED MANPOWER WITH PAPERLESS RECORD:*
Reduce the manpower needed to perform all the record keeping and administration task by reducing the paper works needed.
3) *COST REDUCTION:*
Reduce the cost involved in the student record management process.
4) *OPERATIONAL EFFICIENCY:*
Improve the operational efficiency by improving the quality of the process.

A. Modules Of The Proposed System

The system is designed in such a way that only authorized people are allowed to access some particular modules. The records would be modified by only the administrators. The user would always be in control of the application.

1) *ADMINISTRATOR:*
This module is protected by user ID and password. Regular users of the software will not be permitted to enter into this area of the software. The module will be focusing on the management of users, editing of records and validation and authentication of records.

2) *LECTURER:*
This module is also protected by user ID and password. It consists of all academic staffs that have a hand in students' records. Lecturers will be able to add scores, announcements and events and post grades.

3) *STUDENTS:*
This module is protected by a username and password. This is where students can view anything pertaining to their course such as grades, announcement, and upcoming events.

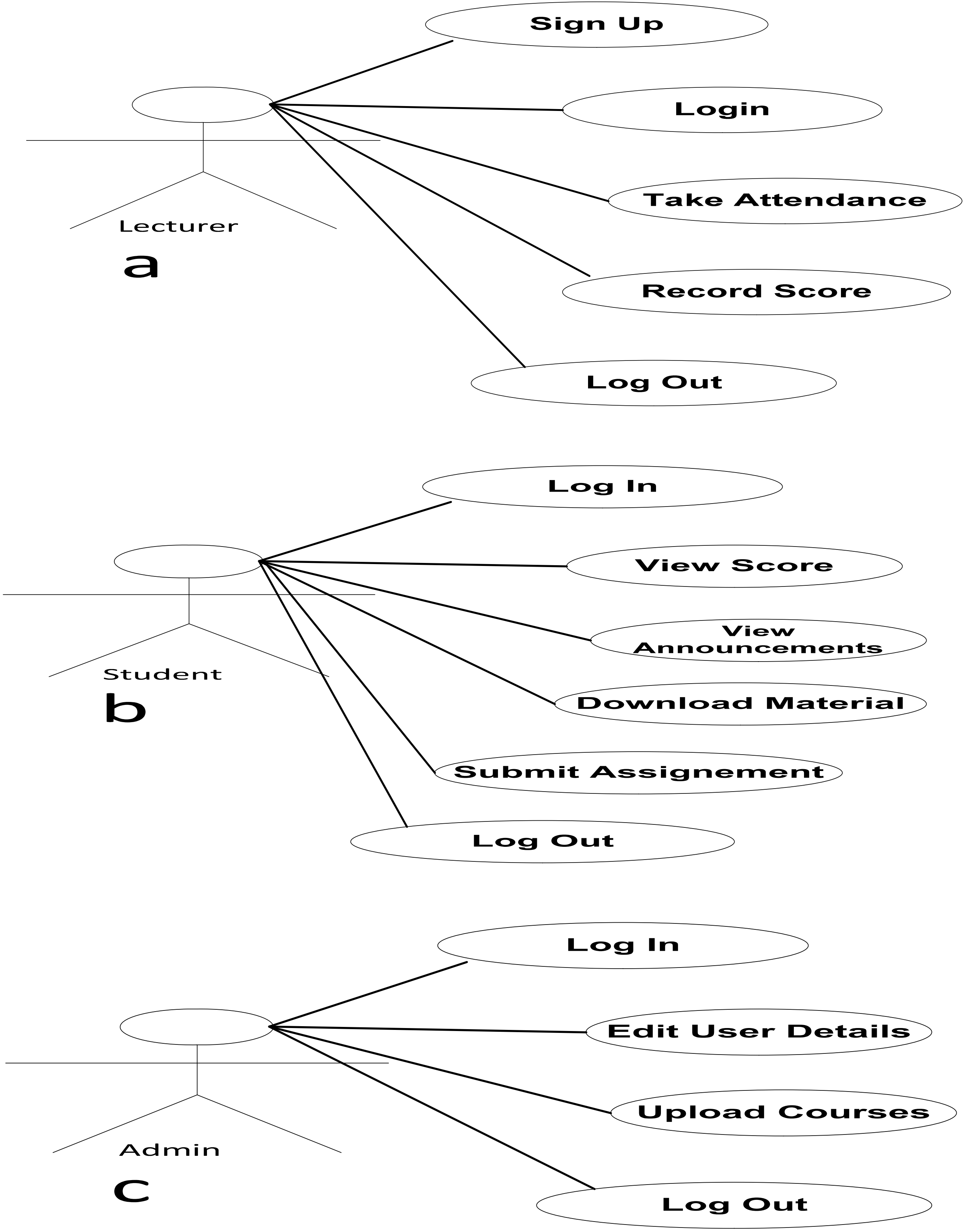


Figure 1: Use case Diagrams

Figure 1 shows the Use Case diagram. This use case has 3 modules; the administrator module, the lecturer and the student module. These modules are discussed below.

B. *Administrator Module:*

This aspect will simply show the function of each entity present in the use case for the Admin**.**

1) *Admin Login Module*
   This is where the administrator is authenticated to access his/her functions in the system using the username and password. If either the username or password is wrong the user is denied access into the system.

2) *Edit User Details Module*
   This is where the administrator edits users' details by either updating their records or deleting their records.

3) *Upload Courses Module*
   This where the administrator uploads the different courses that are available for each semester for students to choose and register to.

C. *Lecturer Module:*

This aspect will simply show the function of each entity present in the use case for the Lecturer**.**

1) *Signup Module*
   This is where the lecturer registers his or herself into the student record management system to be able to use the system.

2) *Lecturer Login Module*
   This is where the lecturer is authenticated to access his/her functions in the system using the username and password. If either the username or password is wrong the lecturer is denied access into the system.

3) *Attendance Module*
   This is where the lecturer takes attendance of students through the system and use these attendance reports to get information regarding the students and evaluate the students' performance.

4) *Record Scores Module*
   This is where the lecturer records students' scores for quiz, assignments, mid semester exams and final exams which can be used to prepare the GPA of students.

D. *Student Module:*

This aspect will simply show the function of each entity present in the use case for the Student**.**

1) *Student Login Module*
   This is where the student is authenticated to access his/her functions in the system using the username and password. If either the username or password is wrong the student is denied access into the system.

2) *View Score Module*
   This is where the student is able to view their scores for their tests, assignments, and exams. This will allow the students know their score beforehand.

3) *View Announcement Module*
   This is where students view announcements from various lecturers concerning the courses they are offering.

4) *Download Materials Module*
   This is where students can download different courses resources put by their lecturers to help them in their studies.

5) *Submit Assignment Module*
   This is where students submit their assignments to their lecturers to be marked and graded and put online for students to view.

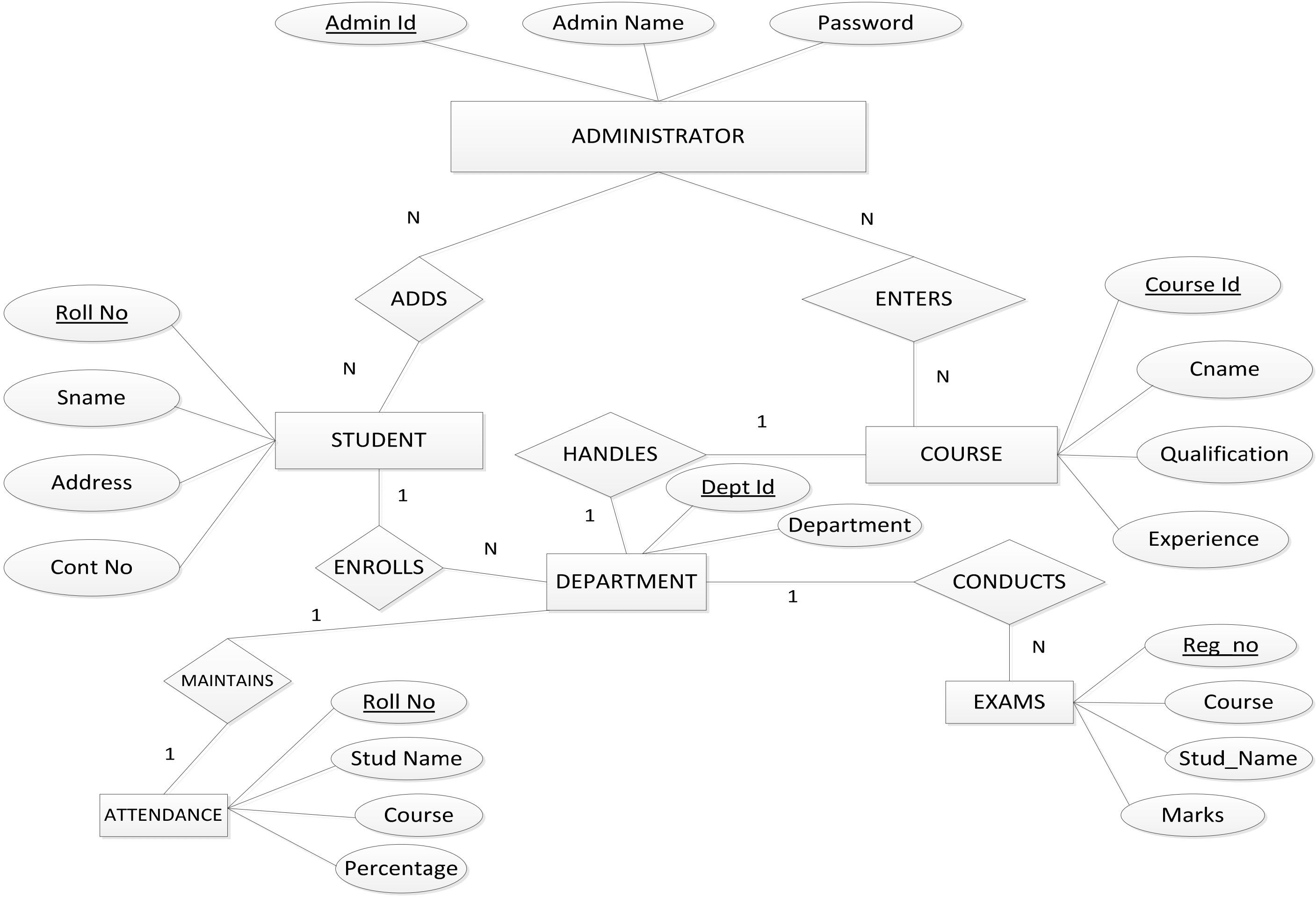


Figure 2: Entity Relationship Diagram for the BSRMS

Figure 2 shows the entity relationship diagram for the BSRMS. This diagram is a graphical representation of entities of the system and their relationships to each other. An entity in this case is a concept about which data is stored while a relationship is how the data is shared between entities. There are 6 entities in this system which are administrator, attendance, exams, department, student and course.

E. Design Tools/Modelling of Proposed System

The functions and benefits of the development tools selected for the design of the proposed system are described in detail.

1) Web Editor

These allows you to create and edit web pages that are either visually or in html codes. A visual editor allows users to create and edit web pages and contents without knowing HTML [16]. Sublime Text Editor was used as the text editor.

2) Hypertext Markup Language

HTML (Hypertext Markup Language) is used to describe how a web page should be displayed. There are various Web browsers and they interpret the HTML codes of Web pages in different ways. When web contents are created, web developers consider the fact that these contents may appear differently when viewed in various browsers [16].

3) Cascading Style Sheet

CSS (Cascading Style Sheets).It defines how to display HTML elements and describes how you want your contents to look like. CSS has rules for specifying the non-visual and visual presentation of web objects/documents. [17]

4) Hypertext Preprocessor

PHP is a scripting language used for the server side of web application development. Example of database supported by PHP are MySQL, Oracle etc.

Why use PHP?
As [18] puts it simply, PHP is the best, fastest and easiest to learn scripting language when it comes to developing dynamic websites.

5) Adobe Photoshop:

This is one of the best photo editors. It was used to create our banners and edit pictures in order to give our site a nice look.

6) MySQL

MySQL is the most popular open-source database system. It is used to develop web based applications. It runs on the server side of applications. Data is stored in what is called tables. A table is a collection of columns and rows [17]. MySQL is very fast, reliable, cheaper and easy to use.

7) *Apache Server*

Apache has been the world's most popular Web server (HTTP server) on the Internet since April 1996 and is generally considered to be more stable than other servers, which is the reason we are making use of it.

8) *Java Script*

JavaScript is a scripting language that is widely supported on most Web browsers. It helps to enhance the user experience on HTML pages. JavaScript is maintained as source code embedded into an HTML page.

F. *System Requirements*

The system requirement contains and describes what the clients want for a particular system. It is a structural document with detailed descriptions of the system services. Some of the system requirements are:

i. Lecturers would have to login to have access to the web application.
ii. Lecturers should be able to view an organized summary of student cumulative score sheet.
iii. Lecturers would be able to take student attendance via biometric device.
iv. Lecturers should be able to upload announcement that can be viewed and downloaded.
v. Students and lecturers should be able to interact through a live chat or forum.
vi. All users should be able to log out.

G. *User Requirements*

Some of the user requirements for this system are:

i. The system requires an internet ready computer
ii. The system will have a database to store all the academic materials that are currently available.
iii. The system requires a biometric authentication device**.**

H. *Functional Requirements*

Some of the functional requirements for this system are:

i. It should take attendance of students.
ii. The system should be able to calculate the grades of students and give the grade point average (GPA) as well as the cumulative grade point average (CGPA).

I. *Non-Functional Requirements*

i. The system will be able to stay up and be active at least 90% of the time. Any downtime or inability to access the system would be due to the system maintenance and upgrade.
ii. The system will have a focused and clear layout. This would help reduce the potential for users to be confused with the interface. It will only display information that is needed by the current task.
iii. The system will have to deal with large quantities of data storage and a large number of users accessing the data at once.

J. *Hardware And Software Requirements*

This section explicitly emphasizes the hardware and software components that is required to run on this application.

K. *Hardware Requirements*

i. An intranet connection
ii. Processor 2.5 GHz
iii. 1014 MB RAM
iv. A biometric device

L. *Software Requirements*

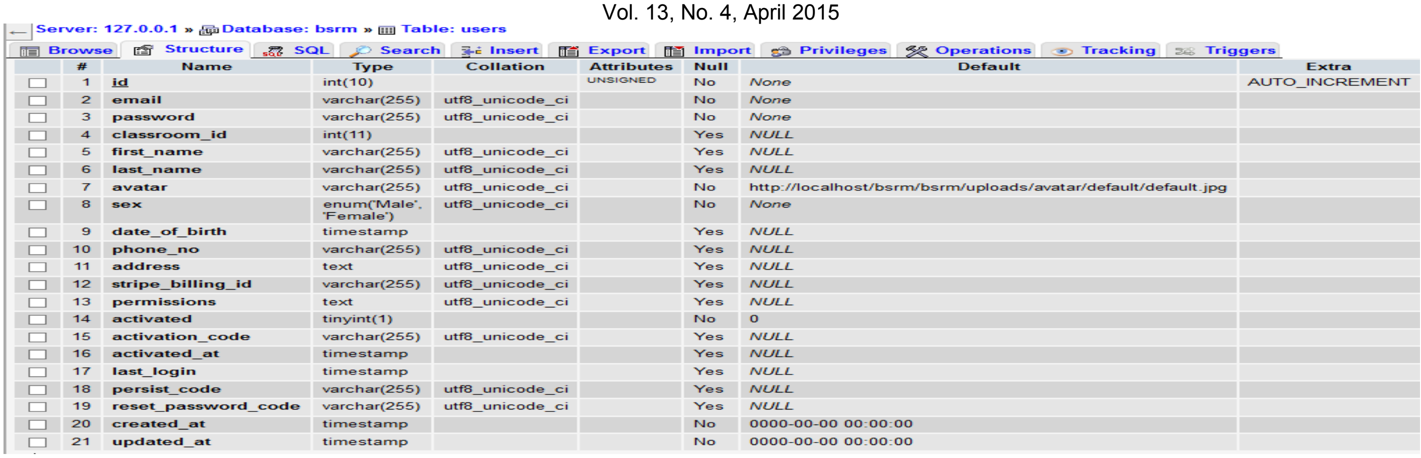
i. Windows Server 2008, Windows 2003 Server or Windows 2000 Server.
ii. SQL Server 2008, SQL Server 2005 or SQL Server 2000 - you can also use SQL Server Express
iii. WAMP server
iv. Operating system of any kind (Microsoft Windows, Mac OS, Linux)
v. Web browser (Safari, Firefox, Internet Explorer, Google Chrome and Opera) must include Java Runtime Environment 1.4 or higher and also JavaScript.

IV. SYSTEM IMPLEMENTATION AND TESTING

This focuses on the implementation and the functional application test for defects and emergent properties such as performance and reliability of the system. A set of tests would be carried out which would involve the execution of the application with test data to ensure that all requirements have been meet correctly in order to ensure high quality and user friendly software. In order to identify potential defects and errors, the software had to go through various types of testing:

A. *Database Testing*

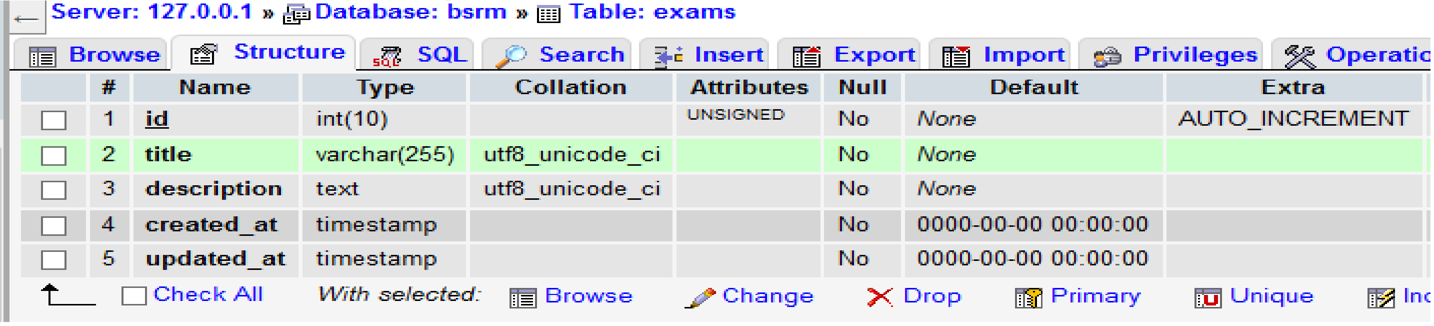
The database is made up of 13 tables and each table contains the name of the fields, data types, sizes, attributes and other constraints that define the table. Below are screenshots of some of the tables along with brief explanations.



	#	Name	Type	Collation	Attributes	Null	Default	Extra
□	1	id	int(10)		UNSIGNED	No	*None*	AUTO_INCREMENT
□	2	email	varchar(255)	utf8_unicode_ci		No	*None*	
□	3	password	varchar(255)	utf8_unicode_ci		No	*None*	
□	4	classroom_id	int(11)			Yes	*NULL*	
□	5	first_name	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	6	last_name	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	7	avatar	varchar(255)	utf8_unicode_ci		No	http://localhost/bsrm/bsrm/uploads/avatar/default/default.jpg	
□	8	sex	enum('Male', 'Female')	utf8_unicode_ci		No	*None*	
□	9	date_of_birth	timestamp			Yes	*NULL*	
□	10	phone_no	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	11	address	text	utf8_unicode_ci		Yes	*NULL*	
□	12	stripe_billing_id	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	13	permissions	text	utf8_unicode_ci		Yes	*NULL*	
□	14	activated	tinyint(1)			No	0	
□	15	activation_code	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	16	activated_at	timestamp			Yes	*NULL*	
□	17	last_login	timestamp			Yes	*NULL*	
□	18	persist_code	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	19	reset_password_code	varchar(255)	utf8_unicode_ci		Yes	*NULL*	
□	20	created_at	timestamp			No	0000-00-00 00:00:00	
□	21	updated_at	timestamp			No	0000-00-00 00:00:00	

Figure 3: users table

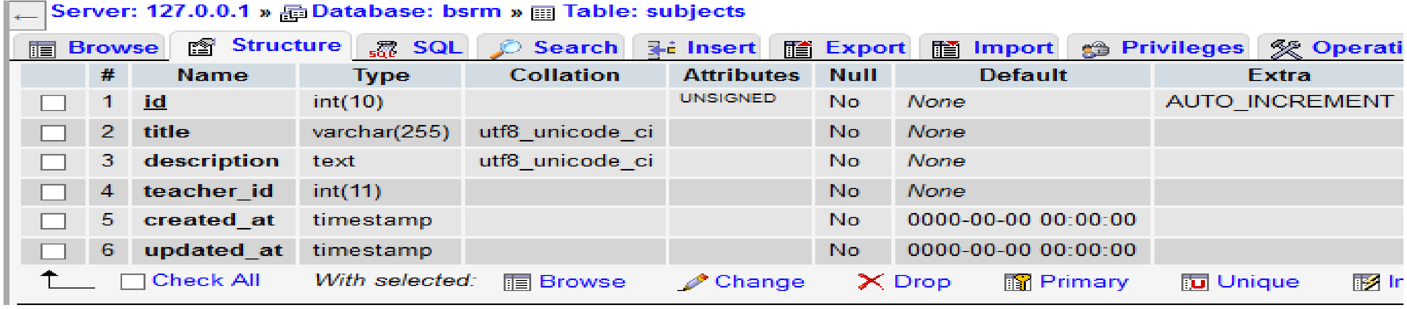
Figure 3 shows the users table which contains the details of the four major users of the system which are the administrator, the lecturers, the students and the parents. This table records the user details such as email address, password, classroom assigned to the user, first name, last name, picture, sex, date of birth, phone no, address, permissions, and information as regards the when each users details were created and updated.



	#	Name	Type	Collation	Attributes	Null	Default	Extra
□	1	id	int(10)		UNSIGNED	No	*None*	AUTO_INCREMENT
□	2	title	varchar(255)	utf8_unicode_ci		No	*None*	
□	3	description	text	utf8_unicode_ci		No	*None*	
□	4	created_at	timestamp			No	0000-00-00 00:00:00	
□	5	updated_at	timestamp			No	0000-00-00 00:00:00	

Figure 4: exams table

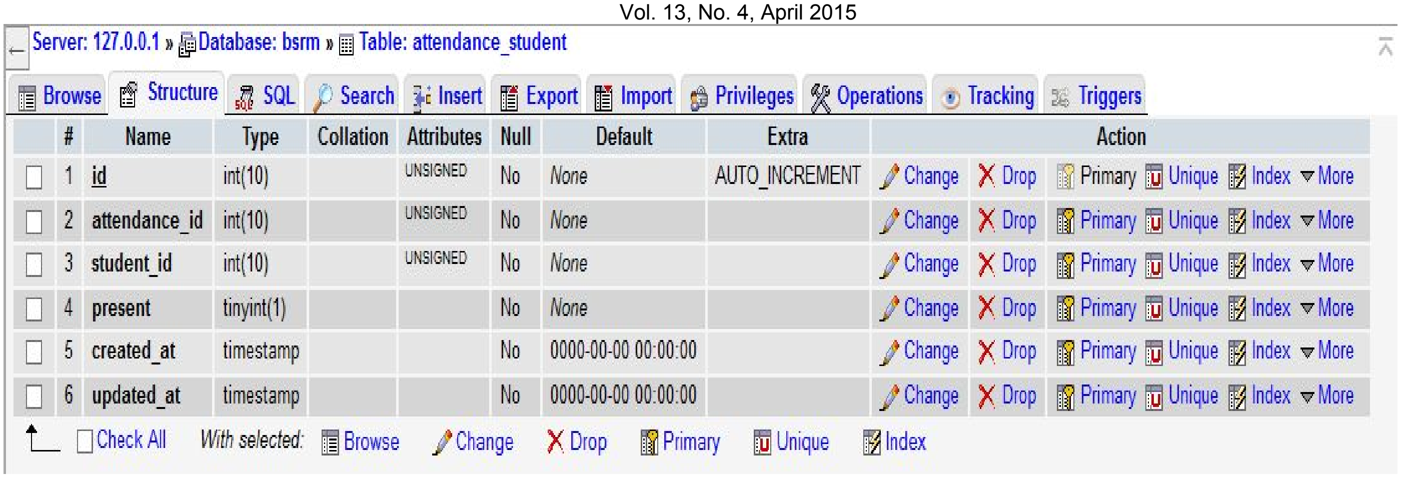
Figure 4 shows the exam table which contains the details of exams scheduled by the administrator. This table records the name of the subject for which the exam is scheduled, the description of what the exam entails and when each exam was created and updated.



	#	Name	Type	Collation	Attributes	Null	Default	Extra
□	1	id	int(10)		UNSIGNED	No	*None*	AUTO_INCREMENT
□	2	title	varchar(255)	utf8_unicode_ci		No	*None*	
□	3	description	text	utf8_unicode_ci		No	*None*	
□	4	teacher_id	int(11)			No	*None*	
□	5	created_at	timestamp			No	0000-00-00 00:00:00	
□	6	updated_at	timestamp			No	0000-00-00 00:00:00	

Figure 5: subjects table

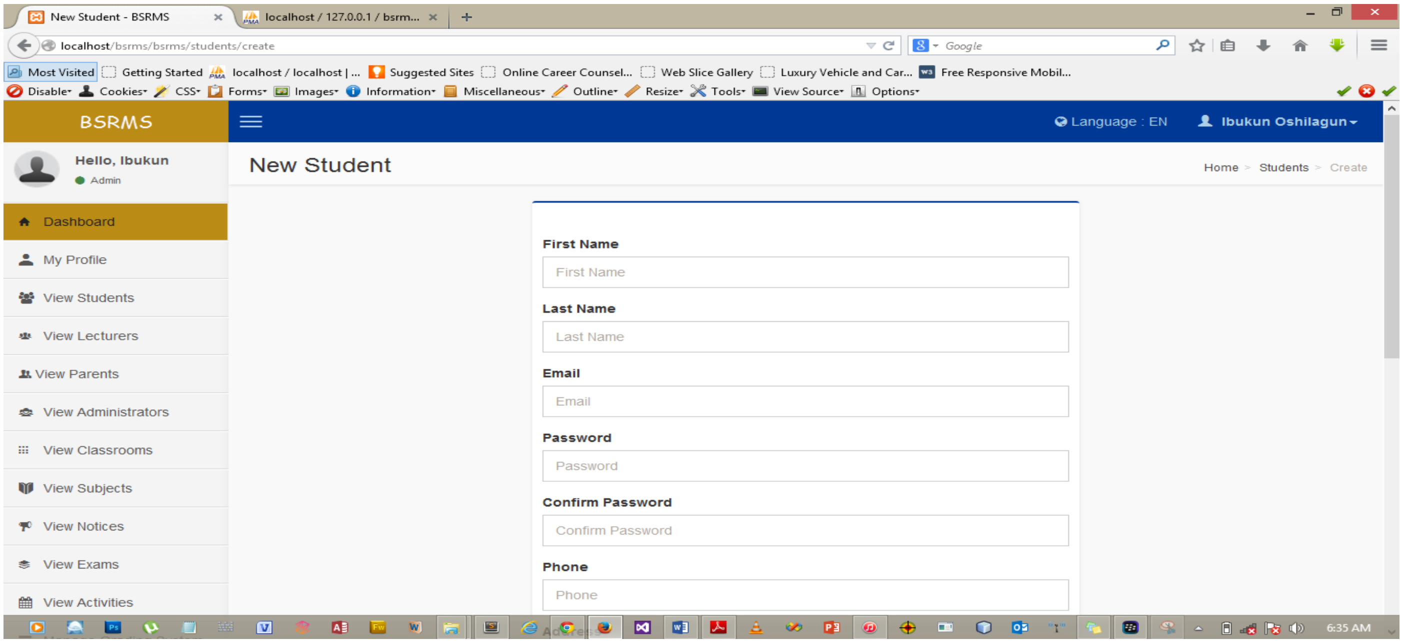
Figure 5 shows the subject table otherwise known as courses table which contains the details of the courses offered in the university. This table records the name of the subject, the description, the name of the teacher assigned to each subject and when information as regards the subject details were created and updated.



	#	Name	Type	Collation	Attributes	Null	Default	Extra	Action
□	1	id	int(10)		UNSIGNED	No	None	AUTO_INCREMENT	Change Drop Primary Unique Index More
□	2	attendance_id	int(10)		UNSIGNED	No	None		Change Drop Primary Unique Index More
□	3	student_id	int(10)		UNSIGNED	No	None		Change Drop Primary Unique Index More
□	4	present	tinyint(1)			No	None		Change Drop Primary Unique Index More
□	5	created_at	timestamp			No	0000-00-00 00:00:00		Change Drop Primary Unique Index More
□	6	updated_at	timestamp			No	0000-00-00 00:00:00		Change Drop Primary Unique Index More

Figure 6: attendance_student table

Figure 6 shows the attendance_student table which contains details about each students' attendance taken during lectures. Each student has an attendance id that is unique to each of them.



New Student

First Name

First Name

Last Name

Last Name

Email

Email

Password

Password

Confirm Password

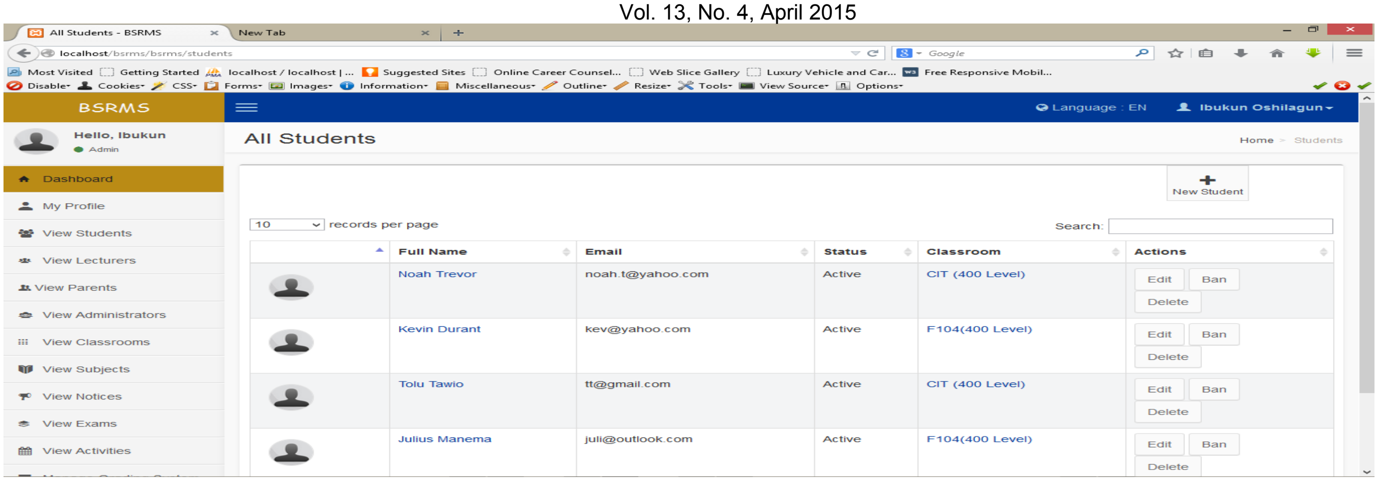
Confirm Password

Phone

Phone

Figure 7: Add new student page

Figure 7 shows the Add New Student page. This page is mainly for the administrator. The administrators are responsible for adding new students into the system. This is done so as to ensure data integrity of the student details being entered into the database. The administrator fills the required fields in the form.

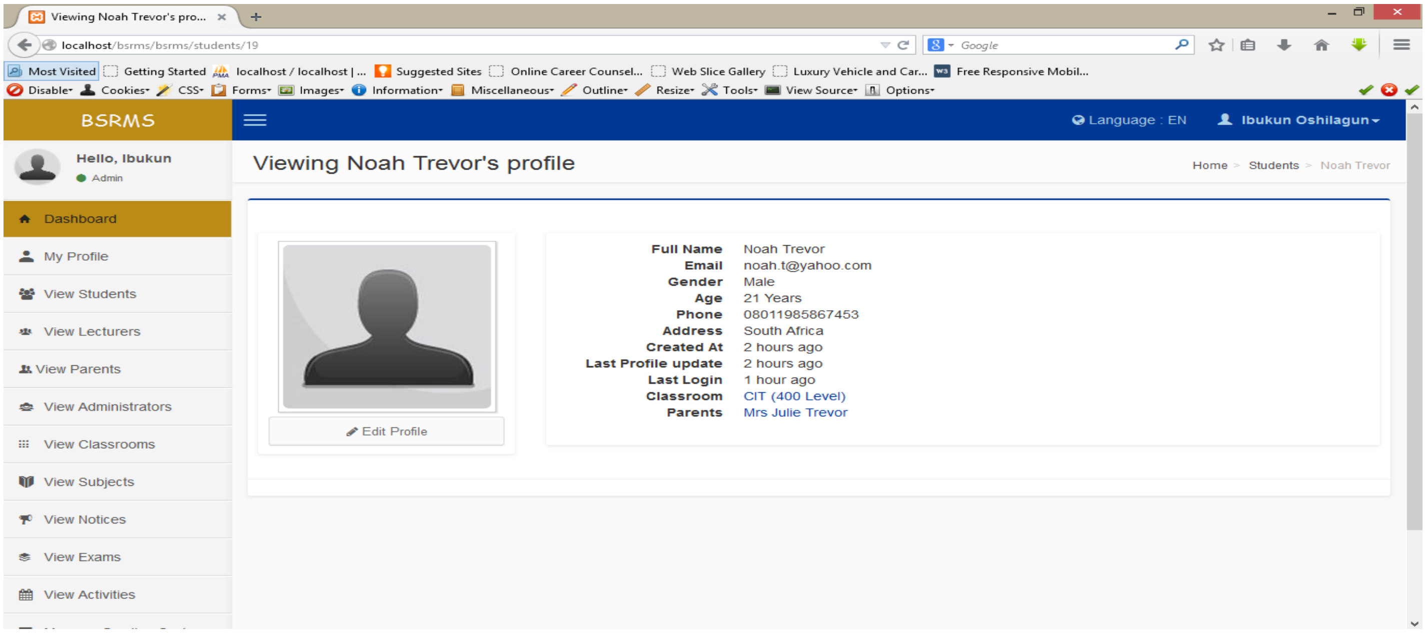


All Students

	Full Name	Email	Status	Classroom	Actions
	Noah Trevor	noah.t@yahoo.com	Active	CIT (400 Level)	Edit Ban Delete
	Kevin Durant	kev@yahoo.com	Active	F104(400 Level)	Edit Ban Delete
	Tolu Tawio	tt@gmail.com	Active	CIT (400 Level)	Edit Ban Delete
	Julius Manema	juli@outlook.com	Active	F104(400 Level)	Edit Ban Delete

Figure 8: view students page

Figure 8 shows the View Students page. This page give a list of all students whose details have been registered and are currently on the system.



Viewing Noah Trevor's profile

Full Name	Noah Trevor
Email	noah.t@yahoo.com
Gender	Male
Age	21 Years
Phone	08011985867453
Address	South Africa
Created At	2 hours ago
Last Profile update	2 hours ago
Last Login	1 hour ago
Classroom	CIT (400 Level)
Parents	Mrs Julie Trevor

Figure 9: view student profile

Figure 9 shows the View Student Profile page. This page displays the profile of each student. The administrators, lecturers and other students on the system have access to this page. Information such as the student's full name, email address, gender, phone number address and last login time are displayed.

B. *Discussion Of Results*

During the course of this research it was realized that developing a biometric student record management system was a herculean task. This project was created to solve the various issues involved with the management of students' records and effectively allows lecturers save time with entering these records. Lecturers with the use of this biometric student record management system will manage students' records more effectively in class and save time. Students will also benefit from this application as they would be able to see announcements online, view timetables and ease the process of taking attendance. To this end, the system developed will improve the way lecturers store and manage students' records and be accountable for students' data and help the university manage the overall records of students more efficiently.

V. SUMMARY

This project work undertook the development of a biometric student record management system application which focused on making the storage and management of students' records efficient to both the school authorities and lecturers with the use of a fingerprint biometric device to ensure the security of these records. This application is web based so that it can run on any platform that supports a working browser. This way it can be assessed and used by anybody with a computer and an internet.

This project was created to solve the various issues involved with the management of students' records and effectively allows lecturers save time with recording these records.

Lecturers with the use of this application would be able to manage students' records more effectively in class and save time. Students would also benefit from this application as they would be able to see announcements online, view timetables and take attendance. To this end, "Biometric Student Record Management System" would improve the way lecturers store and manage students' records and be accountable for students' data and help the university manage the overall records of students efficiently.

A. Recommendations

Pointers to research areas include:

- Students and lecturers should be provided with Internet in order to efficiently use the application.
- Those operating the application must be computer literate. It is therefore recommended that everyone expected to use this application should undergo a computer training programme so as to be proficient in the use of the application.
- Clarification of students who have been registered by the school in order to avoid misunderstanding and complications, this should be done in order to have a correct database.
- For those planning on building this application of this nature in future, a module that creates the administrator must be created instead of creating the administrator from the backend.
- More research should be done in finding ways to make the management of students' records more efficient and safe.
- Other forms of biometric recognition can be used like iris, face or palm.

B. Future Research

Regarding our system, we are planning on introducing more features into the system and possibly implement our system with the current school management system. Also we are also planning on implementing a feature that enables parents check on their kids' activities online.

C. Conclusion

The system successfully took the attendance at lectures and successfully recorded students' details and calculated their grade scores accurately. It also provided a sample time-table created randomly for students. The biometric device successfully captured new fingerprints and stored them in a database. Scanned fingerprints were placed on the sensor and the device compares the current fingerprints with those it has stored in its database. The system performance was good and the system would be considered for full implementation especially because of its short execution time. The system was given out to be tested and results showed that people were pleased and interested in the product being developed for the school.

REFERENCES

[1] Samson.A.A & Adekunbi.A.A. (2013). Design of a Prototype Web-Based Students' Record Management System – WEBSTREMS. Information and Knowledge Management. Vol.3, No.5

[2] U.S. Department of Education, N. C. (2000). Building an Automated Student Record System. Washington, DC 20006: U.S. Department of Education.

[3] Añulika, E. A., Bala, E., & Nyap, C. D. (2014). Design and Implementation of Result Processing System for Public Secondary Schools in Nigeria. International Journal of Computer and Information Technology, Volume 03 – Issue 01.

[4] Bharamagoudar, S.R, Geeta R.B & Totad,S.G (2013). Web Based Student Information Management System. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6.

[5] Adetiba, E., Iortim, O., Olajide, A. & Awoseyin, R. (2013). OBCAMS: An Online Biometrics-based Class Attendance. African Journal of Computing & ICT, Vol 6. No. 3.

[6] Norshidah, K., Helmy, A. W. & Jamal, R. A. (2010). Development of Attendance System using Biometric Fingerprint Identification. Proceedings of EnCon2010 3rd Engineering Conference on Advancement in Mechanical and Manufacturing for Sustainable Environment.

[7] Rishabh.M & Prashant.T. (2011). Student Attendance System Based On Fingerprint Recognition and One-to-Many Matching. Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela-769 008, Orissa, India.

[8] Sulochana.S, Ravindra.T & Balwant.S. (2010). Survey of Biometric Recognition Systems and Their Applications. Journal of Theoretical and Applied Information Technology

[9] Parvathi.A. (2005). Security of Biometric Authentication Systems. 21st Computer Science Seminar. pp1-7.

[10] Njamu O & Ugwu C. (2013). A Novel Web-Based Student Academic Record Information System. West African Journal of industrial and Academic Research. Vol 7, No 1 pp 31- 45.

[11] Satoa, K. & Seema, R. (2013). An Attendance Monitoring System Using Biometrics Authentication. International Journal of Information and Computation Technology, Volume 3, Issue 4.

[12] Idowu, O. & Shoewu, O. (2012). Development of Attendance Management System using Biometrics. Pacific Journal of Science and Technology, 13(1):300-307.

[13] Gunjan .T, Rahul. R & Shete. A.K (2013). Wireless Fingerprint Based College Attendance System Using ZigBee Technology, International Journal of Engineering and Advanced Technology (IJEAT), Volume 2, Issue 3.

[14] Karthik.V.E, Shanmuganathan.S & Sumithra.A. (2013). A Foolproof Biometric Attendance Management System. International Journal of Information and Computation Technology. Volume 3, Number 5 (2013), pp. 433-438

[15] Tutorials point. (2014). sdlc_waterfall_model.htm. Retrieved from tutorialspoint.com http://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm.

[16] UNIVERSITY, C. S. (2010). Web design An Introduction. *Web design An Introduction*, 1,2.

[17] w3c. (n.d.). Retrieved from www.w3cschools.com

[18] Ullman, L. (2008). Visual QuickPro Guide PHP 6 and MySQL 5 for Dynamic Web Sites. Berkeley, United States of America.

AUTHORS PROFILE

ONUIRI ERNEST E. is a doctoral student at the School of Health Related Professions, Rutgers: the State University of New Jersey. He has a Masters in Information Systems at Babcock University. His areas of interest include artificial intelligence, simulation modelling and bioinformatics.

AWODELE OLUDELE is professor of Computer Science (Artificial Intelligence). He is also the chair in the department of Computer Science, Babcock University, Nigeria. He is a fellow, Nigeria Computer Society (NCS) and a member of Computer Professional Registration Council of Nigeria. His areas of interest are Artificial Intelligence, Cloud Computing and Computer Architecture. He has published works in several journals of international repute.

OSHILAGUN IBUKUN is currently a student of Babcock University illishan-remo. He is a final year student studying Computer Information System. His research interest include Web design, photography and music.

ETUK OTOBONG is a student of Babcock University. He is a final year student studying Computer Information System. His research interest include E-Commerce, Web design and development, mobile programming and robotics.

YADI CHUKWUEMEKA is a final year student of Babcock University, Ilishan-Remo, Ogun State studying Computer Information System. He is aspiring to bag a Master’s degree in User experience and also in Human Computer Interaction. His interests are in Web development, User experience, Project Management and Music.

Classification Framework Based on C4.5 Algorithm For Medicinal Data

Karthik Ganesan
Department of Computer Science and Engineering
College of Engineering, Anna University,
Chennai, India

***Abstract*: This study proposes a framework with preprocessing techniques namely Missing value replacement, Discretization, Principal Component Analysis (PCA) to extract the key features and then applying c4.5 classifier algorithm to enhance the classification of medicinal data. The input data gets subjected to missing data imputation through any one of the standard methods like mean, mode, constant and manual input. The dataset is then subjected to Discretization to formalize a reasonable set of discrete bins. PCA is then applied on the dataset to identify the principal components of the dataset, which attribute to the mean data inference. C4.5 algorithm has been used to construct a decision tree based on the information gain of the training set. This work used Cleveland heart disease dataset, obtained from UCI machine learning repository. The dataset is composed of details of about 303 patients and helps to predict presence or absence of cardio vascular disorder based on 75 attributes. The proposed framework was applied on this dataset and exhibited an accuracy of about 77.73%.**

Keywords — PCA, Discretization, C4.5, classification

I. Introduction

Availability of voluminous data in the medical domain has provided an opportunity of extracting useful information from that data. Data mining encompasses techniques such as data cleaning, normalization, classification and association analysis for extracting useful information from data. Furthermore there are mathematical and statistical algorithms have been developed for knowledge mining to arrive at better decision making. The knowledge mined from the data forms the base for computerized medical diagnosis and prognosis systems. The core part of most medical diagnosis and decision support systems consists of a forecasting or prediction system. These are basically supervised machine learning systems trained using real world medical data. The efficiency of these systems, largely rely on the data which is given as input. In order to improve the performance, the data fed into these systems should be pre-processed so that they are clean and balanced with less noise and no outliers.

This work uses a framework that has a data pre-processing module and a classification module. The former primarily deals with assigning values for missing data, equalization of class distribution and dimensionality reduction, whereas the latter employs machine learning tools such as C4.5 classifier algorithm to validate the performance of the processed datasets as well as provide classification for the given input dataset. This work used Cleveland heart disease dataset, a benchmark dataset obtained from the UCI machine learning repository, to predict the presence or absence of cardio vascular disorder. This work discusses various approaches applied in literature for data preprocessing and classification. The study elaborates the need for this framework to improve the efficiency of classification of dataset.

II. LITERATURE SURVEY

A. Data Preprocessing

Yang et al. [10] proposed an adaptive volume data preprocessing approach, which used a semi-automated transfer function for rendering the 3D image for the cardiac MRI data. The system generated high quality 3D images using a median filtering method for data denoising along with adaptive ellipsoidal Gaussian filtering scheme to preserve local features. The preprocessed content provides better clarity of the image. The performance of the system was evaluated using 3D cardio graphic MRI reports consisting of a dataset of size 352 x 352 x 77256. The dataset also consisted of some damaged tissue caused by partial circulation blocking of the heart.

Quing et al. [16] used a Sigmoid function to preprocess the original data and self-organizing neural network to model the output. The Sigmoid function being employed maintains the same geometry of raw data. The system employed four layers namely Integration & Association, Cleaning, Filling & denoising and Normalization to preprocess the raw data. The system was implemented with SOM neuron distribution pattern with a network of 30 x 30, hextop topology and linkdist distance vector function. The system evaluated that training time was directly proportional to the accuracy of data preprocessing and classification with peak of 250, at which the system become convergent.

Zhang et al. [17] proposed a data preprocessing system based on unified data model derived from the analysis of the input data. By using unified dataset as a standard, the system reduced the conventional data transformation / preprocessing of m × n to m + n. The system maintained a permanently stored collection of information containing either case level data or aggregation of case level data. The system also provided flexibility and adaptability for data preprocessing for different data mining tools.

Sabarina et al. [3] proposed a novel technique to adapt the dataset and predict the cancer type. The data was preprocessed through merging and t-testing along with wavelet decomposition. Merging process generated a single data matrix from a number of sample files by taking mean value for missing attributes. The data was subjected to a paired t-test of null hypothesis with 5% significance level to skew the data towards a common direction to achieve better normal distribution of the processed data. The system used SVM for classification purpose, which used a kernel function such as Liner kernel, Quadratic kernel and Gaussian radial basis function kernel. The dataset consists of information pertaining to 128 patients with 12,625 genes as column indices. The performance of the system was evaluated with 2 to 10 fold cross validation and observed to produce high level of accuracy, specificity and sensitivity.

Indre et al. [1] proposed an adaptive preprocessing and predictor component to adapt changes over time. The system implemented an adaptive training system with Baye's learning algorithm to predict the flow of data. They have used Principal Component Analysis (PCA) as the feature extraction tool. In their proposed system, the sensor signals from a chemical production unit was used to adapt the incoming signal from sensor device. The system was designed to identify the principal component and based on that determine whether the reading from a sensor is valid.

The system consisted of a fixed training window which does not require any change detection or online monitoring. This window was periodically used to retrain the predictor using a fixed number of latest historical instances. The system can thus adapt to latest changes in the readings over time. To measure the performance of the system, sensor readings from a chemical production company was used. The dataset consisted of historical data for 3 years with 185 real valued inputs measured every 5 minutes, contributing to a total set of 189,193 instances. The system achieved an accuracy of about 62.17% and 61.46% with Support Vector Machine (SVM) and DT classifier respectively.

B. Classifier

Yi et al. [7] in their work have proposed an approach based on ANFIS for the detection of electro cardio graphic changes in patients with partial epilepsy. The detection of result was arrived based on feature extraction and applying Back propagation algorithm. The wavelet transformation technique was used to extract the features. The detection system was trained with gradient descent back propagation method in combination with the least square method. The systems used eight attributes to process and arrive at the final result of determining the possibility of disease. The processed data was given as input to the system. They created around 108 fuzzy rules in the ANFIS architecture they used. The system was able to detect electro cardio graphic changes due to higher level of accuracy.

Karaolis et al. [11] proposed a data mining system based on decision trees for the assessment of Coronary Heart Disease (CHD) related risk factors targeted in the reduction of CHD events. The events investigated were Myocardial Infarction (MI), Percutaneous Coronary Intervention (PCI) and Coronary Artery Bypass Graft Surgery (CABG). Based on the clinical data obtained from a general hospital, they performed their analysis on 1500 patients during the year 2003-2006 and 2009. They performed analysis using C4.5 decision tree algorithm with five different splitting criteria for extracting rules based on modifiable and non-modifiable risk factors.

The non-modifiable risk factors considered are age, gender, operations, family history and genetic attributes. The modifiable risk factors considered are smoking, hyper tension, diabetes, cholesterol, high-density lipoprotein and triglycerides. They formatted the data and associated a code with them, consisting of a maximum of four codes. They have used the processing techniques like information gain, gini index, likelihood ratio chisquared statistics, gain ratio and distance measure.The highest percentage of classification accuracy achieved by MI, PCI and CABG models were 66%, 75% and 75% respectively.

Eman et al. [4] in their work explained about Coronary Artery Disease (CAD) prediction using an associative classifier that uses Frequent Pattern Growth (FP-Growth). The system generated the set of all frequent class association rules based on the features extracted to predict the highest probability class for the patients. A strong association between the attribute patterns had been searched to generate all rules that satisfy some minimum constraints. Classification rule mining has been used to extract all small set of rules, which combined to build a new classifier. The system was tested with a dataset containing records of 390 patients with abnormal CAD and 280 patients with normal CAD.

III. System Architecture

The proposed system architecture has been depicted in Fig.1. It consists of a preprocessing module and a classifier module. The preprocessing module encompasses missing value imputation component, discretization component and a feature extraction component implemented with PCA algorithm. The classifier module is implemented using C4.5 algorithm to formulate the decision tree based on training dataset and a classifier component iterate the given tuple over the decision tree to obtain the final classification.

A. Preprocessing Module

The preprocessing module is the initial phase of the process wherein the focus is to ensure appropriate data is taken for consideration, thereby producing better outcome. Numerous preprocessing schemes are available, however in this system we consider only missing value imputation and discretization of the data based on supervised method. The preprocessed data has been fed to the PCA component to extract the principal axes.

1) Replace Missing Values

Handling missing values is a data preprocessing technique to obtain a smooth dataset [14]. The common methods include ignoring the tuple that holds missing value, impute with the mean or impute with the most frequent value. In this work, missing values are handled as follows:

- Replacing the attribute value by 0

- Replacing the value by attribute mean
- Update value by getting input from user at run time

The system has provision for both manual and automated way of replacing the missing values. The mean value of an attribute is evaluated and updated in the tuples with empty attribute values. This aids to bring more synergy in the dataset.

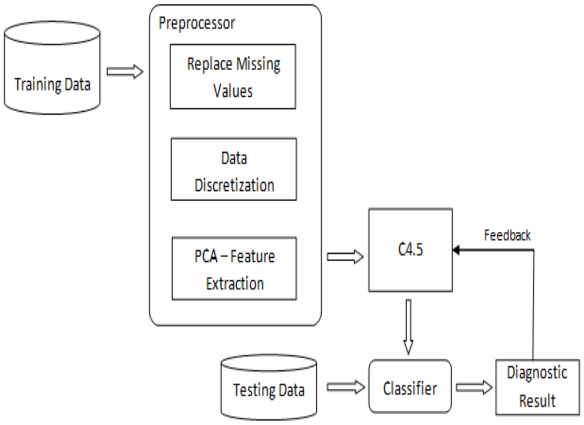


Fig. 1. Proposed system architecture

If the percentage of missing values in a tuple is greater than a defined value, then the corresponding tuple from the dataset will be removed. The system provides facility to select the acceptable percentage of missing values in each tuple. When a tuple does not exhibit any significant correlation, it will be discarded from the set to avoid occurrence of abnormalities.

2) Discretization

The discretization of the dataset has been performed based on chi-square statistic method. The continuous attributes like Age can be discretized to form specific bins, which represent a range of values. The binning of attribute values help in reduction of major outliers and accommodate the values in appropriate group.

Input

Continuous attribute data

Process

Step 1: Select the attribute with continuous data

Step 2: Sort the data values in ascending order

Step 3: Calculate the Chi-square statistic

Step 4: Identify the set of bins to discretize

Step 5: Enumerate the attribute values and update with appropriate bin

Output

Discrete bins

3) Principal Component Analysis

In order to handle multi attribute nature of the data, PCA [1] is used. Through dimensionality reduction technique, the input data is converted from multidimensional input to multidimensional output. The pre-processor thus maps the input data to a set of output data aligned to a principal axis. It helps in extracting the feature vector from the available set of input data. A rotation matrix is learned based on the input data and fixed. Though the data is rotated using PCA, the feature will be preserved. It provides consistent results and robust though a feature with variable attribute gets added to the system.

Input

Multivalued attribute data

Process

Step 1: Calculate the empirical mean of the given dataset

$$u[j] = 1/n \sum X(i, j) \quad (2)$$

Step 2: Find the covariance matrix

$$C = 1 / (n-1)\ B^T \times B \quad (3)$$

Step 3: Identify eigen vectors and eigen values

$$V^{-1} \times C \times V = D \quad (4)$$

Step 4: Rearrange the eigen vectors

Step 5: Obtain the principal component

Output

Linear data with uncorrelated output

4) Decision Tree Generation using C4.5

C4.5 is one of the statistical classification algorithms which generate decision trees based on information gain of the training dataset. The training tuples must contain a set of attributes along with class attribute. The algorithm supports both continuous and discrete attribute values. The data can be of either numeric or nominal value. The class attribute must be a discrete value, which maps a set of attribute values to a particular class.

Input

Training tuples with class attribute

Process

Step 1: Create a root node N

Step 2: If tuples are of same class, then return the class name

Step 3: Return root node with the class labelled

Step 4: If attribute list is empty, then return root node with majority of the class value

Step 5: Calculate the frequency of each sub class C in the set S using (5)

$$p_i = \text{frequency } (C_i, S) / |S| \quad (5)$$

Step 6: If the attribute list contains varying class values in the dataset S, then calculate information gain of the tuple using (6)

$$\text{Info } (S) = \sum - p_i \log_2 p_i \quad (6)$$

Step 7: Extract the possible set of attributes with their values sorted by ascending order

Step 8: Calculate the information gain for each subset of attributes, which can be split further

$$\text{Info}_A(S) = \sum (|S_j| / |S|) \times \text{Info } (S) \tag{7}$$

Step 9: Calculate the gain of each attribute based on the attribute of the node

$$\text{Gain } (A) = \text{Info } (S) - \text{Info}_A(S) \tag{8}$$

Step 10: Select the attribute A with maximum gain

Step 11: Create a node with attribute A and link it to its parent node N

Step 12: If the nodes created for a particular parent reaches the maximum threshold set, then return as leaf node with maximum class value selected

Step 13: If the tuples obtained by splitting on the attribute A have same class value, then return as a leaf node

Step 14: If tuples obtained have different class values, then create a decision node and return to Step 2

Step 15: Return the decision tree

Output
Decision tree

IV. Experimental Results And Discussion

The system was experimented with Cleveland heart disease dataset from UCI machine learning repository. The dataset has 13 attributes with class label. The class label classifies the sample as presence or absence of heart disease. Class label '1' indicates presence of heart disease and class label '0' indicates absence of heart disease. Classes '2', '3' and '4' represent the severity of disease and the last two have been merged in this experiment as part of approximation. The dataset has 303 tuples with no missing values, out of which 100 tuples are used for training and the system was experimented with remaining tuples. The results of the experiments are tabulated in Table I.

TABLE I. Confusion Matrix Of Classification Results

Expected	**Observed**			
	0	***1***	***2***	***3***
0	83	20	4	0
1	6	25	5	0
2	0	4	18	4
3	0	4	8	22

The confusion matrix consists of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values. Accuracy, sensitivity and specificity are computed based on the confusion matrix.

True Positive Rate (TPR) and False Positive Rate (FPR) are also other performance metrics. TPR can be related to sensitivity and hence FPR becomes 1 – Specificity.

Accuracy (9) is the percentage of the sample data that are correctly classified by the classifier.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \tag{9}$$

Sensitivity is the percentage of how many samples are correctly classified as true positive.

$$\text{Sensitivity} = TP / (TP + FN) \tag{10}$$

Specificity is the percentage of how many true negatives is predicted by the classifier.

$$\text{Specificity} = TN / (TN + FP) \tag{11}$$

Precision is the percentage of true positive amongst the overall positives predicted by the classifier.

$$\text{Precision} = TP / (TP + FP) \tag{12}$$

TABLE II. Evaluation Measure Results

Evaluation Measure	***Result %***
Accuracy	77.3
Sensitivity	68.2
Specificity	87.5
Precision	85.8

The C4.5 classifier has generated a decision tree based on the training dataset. The testing dataset applied on the decision tree has resulted in the classification as mentioned in Fig. 2

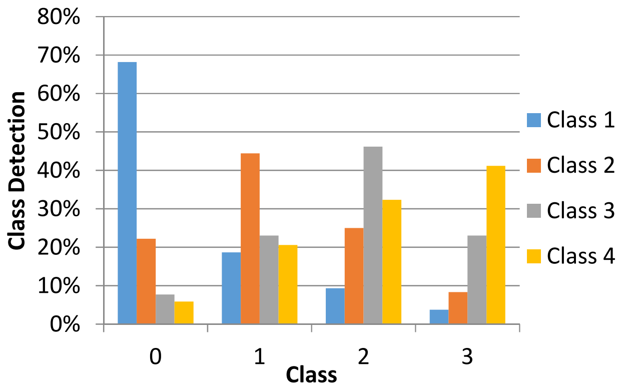


Fig. 2. Experimental results of classification

V. Conclusion And Future Enhancements

In this work, a framework has been designed for classifying the Heart disease data. The input data is preprocessed by replacing missing values, discretizing the continuous variables and feature extraction using PCA. A decision tree has been generated using C4.5 algorithm. The system has achieved an accuracy of 77.3% to classify the dataset. The proposed framework has controlled the creation of branches beyond a threshold, thereby limiting the creation of unnecessary branches.

This work can be extended further by improving the discretization algorithm to handle attribute values with larger range. An approach to further reduce the tree construction time can be investigated.

REFERENCES

[1] Indre Zliobaite and Bogdan Gabrys, "Adaptive Preprocessing for Streaming Data", IEEE Transactions on Knowledge and Data Engineering, vol 26, No. 2, pp. 309-321, February 2014.

[2] Rahul R, John Rushing, Amy Lin, Helen Conover, Xiang Li and Saras Graves, "Data Prospecting: An Approach Towards Data Intensive Science", IEEE Journal of Selected Topics In Applied Observations and Remote Sensing, vol 6, No. 3, pp. 1233 – 1241, June 2013.

[3] Sabarina Rashid and Golam Morshed Maruf, " An adaptive feature reduction algorithm for cancer classification using wavelet decomposition of serum proteomic and DNA microarray data", IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 305 – 312, 2011.

[4] Eman AbuKhousa and Piers Campbell, "Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", International Conference on Innovations In Information Technology, pp. 267 – 272, 2012.

[5] Zhang N. And W.F.Lu., "An Efficient Data Preprocessing Method for Mining Customer Survey Data", 5th IEEE International Conference on Industrial Informatics, Vol. 1, pp. 573 – 578, 2007.

[6] Michele Berlingerio, Francesco Bonchi, and Fraco Turini," Mining Clinical Data with a Temporal Dimension: A Case Study", IEEE International Conference on Bio Informatics and Biomedicine, pp. 429 – 436, 2007.

[7] Yi Mao, Yixin Chen, Gregory H. And Minmin Chen, "Medical Data Mining for Early Deterioration Warning in General Hospital Wards", IEEE International Conference on Data Mining Workshops, pp. 1042 – 1049, 2013.

[8] Aboulnasr Hassanien and Sergiy A. Vorobyov, "New Results On Robust Adaptive Beamspace Preprocessing", Sensor Array and Multichannel Signal Processing Workshop, pp. 315 – 319, Jul 2008.

[9] Jun Zheng, Ming Zeng Hu and Hong Li Zhang, "A New Method of Data Preprocessing And Anomaly Detection", Proceedings of the Third International Conference on Machine Learning and Cybernetics, pp. 2685 - 2690, August 2004.

[10] F Yang, WM Zuo, KQ Wang and H Zhang, "3D Cardiac MRI Data Visualization Based on Volume Data Preprocessing and Transfer Function Design", IEEE – Computers In Cardiology, pp. 717 – 720, September 2008.

[11] Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., &Pattichis, C. S., "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees", IEEE Transactions on Information Technology in Biomedicine, vol. 14, pp. 559-566, 2010.

[12] Ricardo Bellazzi, Lucia Sacchi, and Stefano Concaro, "Methods and Tools for Mining Multivariate Temporal Data in Clinical and Biomedical Applications", IEEE International Conference of IEEE EMBS, pp. 5629 – 5632, 2009.

[13] Kai Liu, Xuezhong Zhou, Yan Feng and Jie Liu, "Clinical Data Preprocessing and Case Studies of POMDP for TCM Treatment Knowledge Discovery", IEEE 14th International Conference on e-Health Networking, Applications and Services, pp. 10 – 14, 2012.

[14] Jiawei Han & Michellin Kamber, "Data Mining, Southeast Asia Edition: Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[15] Sudheer Reddy, Kantha Reddy and Sitaramulu, "An effective Data Preprocessing method for Web Usage Mining", International Conference on Information Communication and Embedded Systems, pp. 7 – 10, 2013.

[16] Qing Ang, Weidong Wang, Zhiwen Liu and Kaiyun Li, "Explored Research on Data Preprocessing and Mining Technology for Clinical Data Applications", The 2nd IEEE International Conference on Information Management and Engineering (ICIME), pp. 327 – 330, 2010.

Energy Efficiency of IEEE 802.15.6 MAC Access Modes for Remote Patient Monitoring Applications

[1]Anas Bouayad, [2]Nour El Houda Chaoui, [3]Mohammed El Ghazi, [4]Molhime El Bekkali
[1234]TTI Laboratory
USMBA
FEZ, MOROCCO

Abstract- **The progress that has been made over the last decade in the medical field was focused on integrating communication and information technology especially Wireless Body Area Networks (WBANs) in healthcare systems for remote patient monitoring (RPM) applications. WBANs have shown great potential in improving healthcare quality, allowing continuous patient to be remotely monitored and diagnosed by doctors. WBAN operates in close vicinity to, on, or inside a human body and supports a variety of medical applications. Energy consumption is a key WBANs since energy-constrained sensors monitor the vital signs of human beings in healthcare applications. In this work, we are interested in evaluating access methods and 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 decrease the energy consumption. 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, (MAC) protocols, access methods, polling, CSMA/CA, Energy consumption.

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 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. Wireless technologies is one of these tools.

The increase use of wireless networks and the constant miniaturization of electrical devices have empowered the development of Wireless Body Area Networks (WBANs). 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.

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 802.15.6 is considered as a promising standard 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.

Within this context, the objective of this work is to model and simulate a heterogeneous WBAN allowing the measurement and the transmission of short range data collected by the environmental sensors. 11 nodes, seems to be sufficient for this monitoring application. These nodes will exchange data between them according to a communication protocol that optimizes energy consumption, transmission delay and loss of information.

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, and presents an energy consumption analysis. Section V 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 VI.

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 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 in terms of energy consumption still few compared to other standard. For example: An analysis was developed for packet size optimization under m-periodic scheduled access mode, to improve the efficiency of energy consumption [14]. In [15] we find an energy analysis that investigates the impacts of scheduled access of an IEEE 802.15.6 for medical applications.

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).

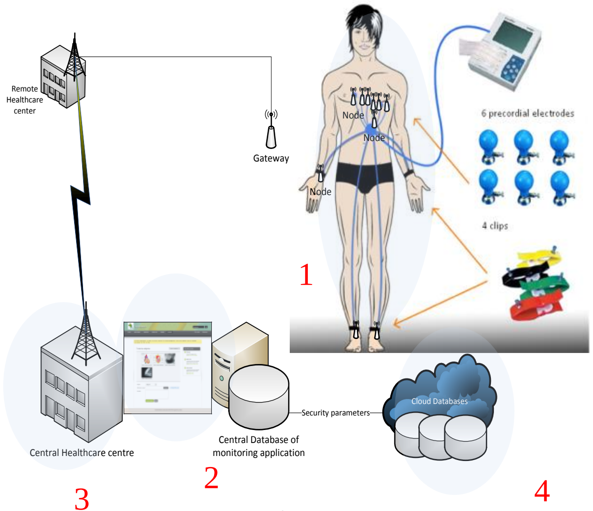


Fig.1 Architecture of the solution

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). 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 polling and CSMA/CA mechanisms.

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 three categories: random 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 NB PHY is chosen.
- Slotted Aloha in case of choosing UWB PHY.

In this paper we are interested to the narrowband (NB) Physical layer.

2) Improvised and unscheduled access

- 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 MAC 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.

- 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

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.

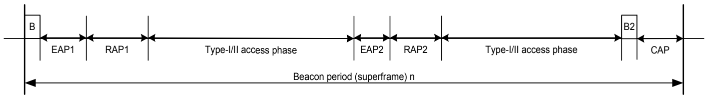


Fig. 2 Layout of access phases in a beacon period (superframe) for beacon mode

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 or hub). 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 coordinator 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 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.

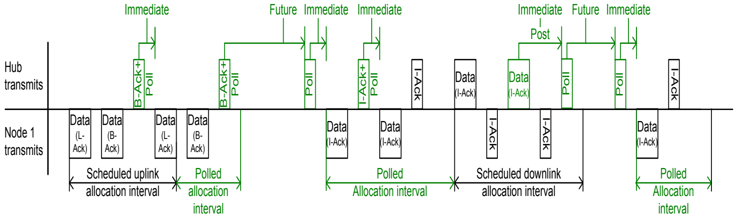


Fig. 3 Example of polled allocation [16]

C. *Sleep mode and energy saving*

Energy efficiency is increased via mechanisms that allow sensor nodes to enter a low-power sleep mode for a long time (several beacon periods) before transmitting / receiving. The sleep mode is explained as follows. 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.

D. CSMA/CA mechanism

This method of access uses a backoff counter and a contention window in order to obtain a new contended allocation. A node can start, use, modify, abort and end a contended allocation. The CSMA/CA mechanism defined in the IEEE 802.15.6 standard is shown in Fig. 4 and its procedure is explained as follows. A node initialize its counter to a random integer value between the interval [1, CW] where CW Є [CW_{min}, CW_{max}], depending to the user priority (UP) as shown in table 1. If the CW is small, the case of emergency traffic, there is a high probability to access the channel. If the CW is large, the case of regular traffic, there is a low probability to access the channel. Each sensor decrements the backoff counter by one for every idle CSMA slot. When the backoff counter value of a node reaches zero, then the node transmits a packet, and the CW is configured as follows: It is set to CW_{min}, if the node did not obtain any contended allocation or if the frame transmission was successful. The CW is not changed, if the transmitter node does not require an ACK frame or if this is its m^{th} time where the node has failed consecutively, with m being odd number. The CW is doubled if the node has failed consecutively n (even) times. If after doubling CW, it exceeds CW_{max}, the CW is set to CW_{min}. The backoff counter is locked by the node until the end of the current frame transmission if the channel is busy or if the backoff counter is reset. It is also locked if the current time is outside of RAP and CAP for regular traffic or if the current time is outside of EAP, RAP, and CAP for emergency traffic. Moreover, when there is not enough time to finish the current transmission, the backoff counter is also blocked. On the other hand, the backoff counter is unlocked when the channel is idle for the pSIFS period within a CAP or RAP for regular traffic and when there is enough time to finish the current transmission [21].

Table 1
Contention window (CW) bounds and UP mapping for CSMA/CA.

Priority	**UP**	**Traffic designation**	**CW_{min}**	**CW_{max}**
Lowest	**0**	Background (BK)	16	64
	1	Best effort (BE)	16	32
	2	Excellent effort (EE)	8	32
	3	Controlled load (CL)	8	16
	4	Video (VI)	4	16
	5	Voice (VO)	4	8
	6	Media data or network control	2	8
Highest	**7**	Emergency or medical event report	1	4

An example of the CSMA/CA procedure defined in the IEEE 802.15.6 standard is shown in Fig.4 [21].

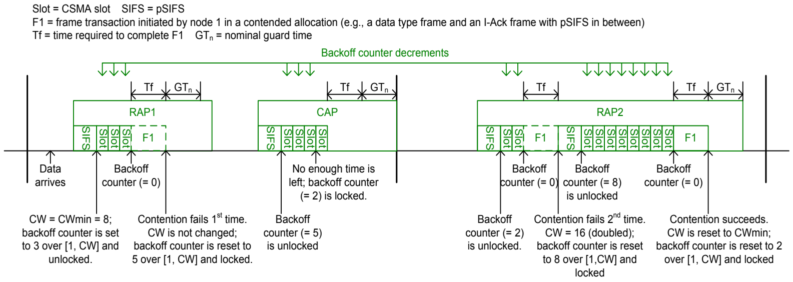

Slot = CSMA slot SIFS = pSIFS
F1 = frame transaction initiated by node 1 in a contended allocation (e.g., a data type frame and an I-Ack frame with pSIFS in between)
Tf = time required to complete F1 GTn = nominal guard time
Backoff counter decrements
Tf
GTn
RAP1
CAP
RAP2
SIFS
Slot
F1
Data arrives
Backoff counter (= 0)
No enough time is left; backoff counter (= 2) is locked.
Backoff counter (= 0)
Backoff counter (= 8) is unlocked
Backoff counter (= 0)
CW = CWmin = 8; backoff counter is set to 3 over [1, CW] and unlocked.
Contention fails 1st time. CW is not changed; backoff counter is reset to 5 over [1, CW] and locked.
Backoff counter (= 5) is unlocked
Backoff counter (= 2) is unlocked.
Contention fails 2nd time. CW = 16 (doubled); backoff counter is reset to 8 over [1,CW] and locked
Contention succeeds. CW is reset to CWmin; backoff counter is reset to 2 over [1, CW] and locked

Fig. 4 Example of IEEE802.15.6 CSMA/CA mechanism

E. Analysis

1) Transmission Time

Several parameters are defined as follows:
T_{t} : Transmission Time
$T_{CW:}$ backoff time
T_D : Time to transmit a data packet
T_{pSIFS} : Interframe spacing
T_{ACK} : Time of packet acknowledgement
T_{cs} : CSMA slot length
T_{ε} : delay time
T_P : time to transmit a preamble
T_{PHY} : time to transmit physical header
T_{MAC} : time to transmit MAC header
T_{BODY} : time to transmit MAC frame body
T_{FCS} : time to transmit frame check sequence
R_s : Symbole Rate
R_{data} : Data Rate
PL : Paylod Size
PL_b : Paylod Size for block acknowledgement

The T_t is defined as total time to transmit a data packet included the T_{CW}, T_D, T_{pSIFS}, T_{ACK} and T_{ε}.

$$T_t = T_D + T_{CW} + T_{ACK} + 2T_{pSIFS} + 2T_{\varepsilon} \quad (1)$$

The average backoff time can be obtained as follows:

$$T_{CW} = \frac{CW_{min}.T_{CS}}{2} \qquad (2)$$

Since a data packet consists of a preamble, physical header, MAC header, MAC frame body and frame check sequence, the transmission time of a data packet becomes as:

$$T_D = T_P + T_{PHY} + T_{MAC} + T_{BODY} + T_{FCS} \qquad (3)$$

$$= \frac{\text{Preamble + PHY header}}{R_S} + \frac{\text{8.(MAC header+PL+ MAC footer)}}{R_{data}}$$

Since an immediate acknowledgement carries no payload, its transmission time is given by:

$$T_{I\text{-}ACK} = T_P + T_{PHY} + T_{MAC} + T_{FCS} \qquad (4)$$

$$= \frac{\text{Preamble + PHY header}}{R_S} + \frac{\text{8.(MAC header+ MAC footer)}}{R_{data}}$$

$$T_{B\text{-}ACK} = T_P + T_{PHY} + T_{MAC} + T_{BODY} + T_{FCS} \qquad (5)$$

$$= \frac{\text{Preamble + PHY header}}{R_S} + \frac{\text{8.(MAC header+PLb+ MAC footer)}}{R_{data}}$$

2) *Energy consumption*

We are interested in analyzing the energy consumed when a sensor node *n* in a WBAN sends data towards the gateway *g* using single-hop communication. The energy consumed is:

$$E_{comm} = E_{TX} + E_{RX} \qquad (6\ )$$

Where E_{TX} is the energy consumed by the transmitter (node *n*), and E_{RX} is the energy consumed by the receiver (node *g*).
T_{SB}: backoff time for a packet successfully transmitted
T_C : collision time,
T_D : time to transmit a data packet,
T_{Dp} : time for a packet to be dropped,
T_{DpB} : backoff time for a packet dropped,
N_C : number of collisions for a packet successfully transmitted,
E_{Dp} : energy consumed by a node due to a packet dropped,
E_{DpB} : energy consumed during backoff for a packet dropped,
E_{DpC} : energy consumed due to collision for a packet dropped,
E_S: energy consumed during successful transmission of packet,
E_{SB} : energy consumed during backoff for a successful transmission,
E_{SC} : energy consumed during collision for a successful transmission,
E_{TX} : energy consumed when transmitting a packet,

For a successful packet transmission, the consumed energy is given as:

$$E_S = E_{TX} + E_{SB} + E_{SC} \qquad (7)$$

Where E_{TX} is given as :

$E_{TX} = E_{TX,\,D} + E_{RX,\,A}$

D and A refer to DATA and ACK packets, respectively.

$E_{TX,\,D} = T_{TX} \,.\, P_{TX}$
$E_{TX,\,A} = T_{ACK} \,.\, P_{RX}$

E_{SB} and E_{SC} are given as :

$E_{SB} = T_{SB} \,.\, P_{IDLE}$

$E_{SC} = N_C \,.\, (\, P_{TX} \,.\, T_{TX} + T_C + P_{TX} \,.\, T_{TX})$

For an unsuccessful packet transmission, the consumed energy is given as:
$$E_{Dp} = E_{DpB} + E_{DpC} \qquad (8)$$

Where E_{DpB} and E_{DpC} are given as :

$E_{DpB} = P_{IDLE} \,.\, T_{DpB}$

$E_{DpC} = (r\text{-}1).\,(\, P_{TX}.T_{TX} + T_C + P_{TX}.T_{TX})$

3) *Device life time*

The device lifetime can be estimated from the total current consumption (I_{total}) and the battery capacity (Q_b).

$$T_{life} = \frac{Q_b}{I_{total}} \qquad (9)$$

I_{total} is calculated as follows :

$I_{total} = I_{Rx} + I_{Tx} + I_{IDLE} + I_{transition} + I_{Sleep}$

Where I_{Rx} and I_{Tx} are the current consumption when the device are in transmission or receive state, I_{IDLE} is the current consumption for Idle state, $I_{transition}$ is the current consumption when a node change the state from idle to transmitting or receiving, I_{Sleep} is the current consumption for a node in the sleep mode.

The current consumption in each state described above, can be calculated using the formula bellow [18]:

$$I_{state} = \frac{T_{state}.\, i_{state}}{m.\, T_{SF}} \tag{10}$$

From (9) and (10), the device life time can be expressed as:

$$T_{life} = \frac{m.\, Q.\, T_{SF}}{T_{state}.\, \mathrm{i}_{state}} \tag{11}$$

Where the superframe duration is:
T_{SF} = nSlots + T_{slot}
T_{slot} =(pAllocationSlotMin + L.pAllocationSlotResolution)

$$T_{life} = \frac{m.\, Q.\, (T_{slot} \ + nSlots)}{T_{state}.\, \mathrm{i}_{state}} \tag{12}$$

Form (12), it can be concluded that the parameters nSlots, which can be between 1 and 256 and the slot length L, which can be between 0 and 255, influence the device life time.

V. SIMULATION PARAMETERS AND RESULTS

A. Simulation Parameters

The simulation framework we choose is the Castalia open source simulator [19]. All simulations described in this paper are released with Castalia 3.2, assisting with the reproducibility of the results. Fig.5 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.

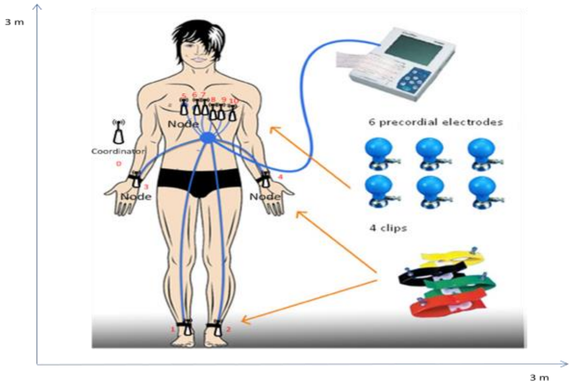


Fig.5 Simulated Network Topology

The radio parameters we define that meet with the IEEE 802.15.6 radio proposal [20] 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 2 gives the various radios parameters defined.

Table 2
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

The effect of path loss is considered from [21]. 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 [22].

Table 3
MAC default parameters defined

Slot allocation length	10 ms
Allocations slots in a beacon period	32 slots
Requesting slots per node	0 slots
Contention based access slots	8
Buffer MAC	48 packets
Retransmission packets tries	2
Polling mechanism	Enabled
Packets Rate	30

Table 3 give the most important default parameters used in the simulation scenario.

The initial energy budget of a node is 18720 joules.

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.

B. Simulation Results

We observed the performance of different access mechanisms of the MAC protocol in terms of energy consumption.

The energy consumptions of the protocol IEEE 802.15.6 is compared for different access mechanisms, as presented in the section IV. The results of energy consumption presented are averages on all nodes, for a fixed packet rate.

1) *Energy consumption when random access mode is used*

The consumed energy graph presented bellow (Fig.6) shows the average consumed energy per node for different length of possible random access slots. The packet rate is fixed on 30pkt/secs/node. The results are presented for two cases: polling mechanism enabled and polling disabled.

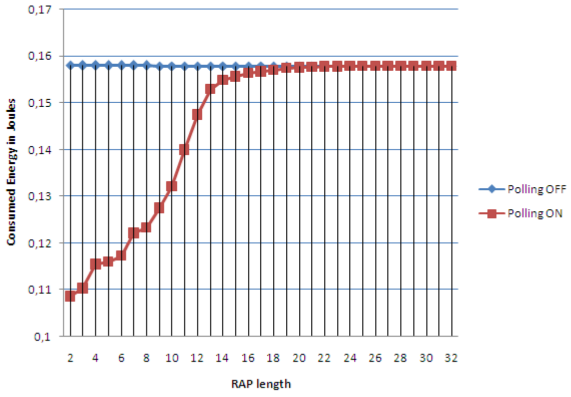


Fig. 6 Average Energy Consumption for different length of RAP

As shown in the graph, the first characteristic we notice is the increase of consumed energy as the duration of random access increases, better performance (less consumed energy) is obtained when the number of random access length decrease.

The simple rationale is that as we have more contention time, we have more opportunities for packets to collide (since CSMA is imperfect). This contention leads to use more energy when nodes are contending for long periods of time, since the energy consumption, basically reflects how long the radio is staying in the listening or receiving mode.

This is why 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 and collision.

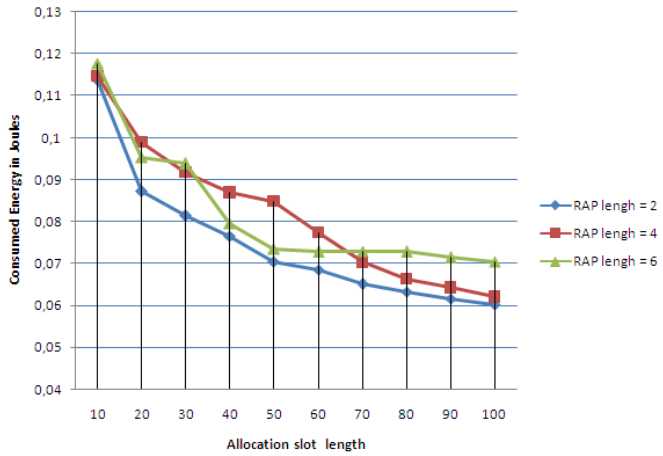


Fig. 7(a) Average Energy Consumption for different length of allocation slot (beacon period = 32 slot)

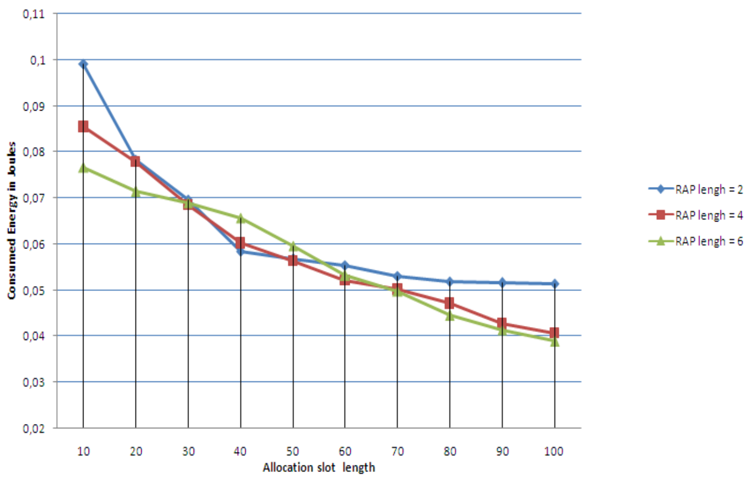


Fig. 7 (b) Average Energy Consumption for different length of allocation slot (beacon period = 128 slot)

In the graphs above (Fig.7a and Fig.7b), we observe clear reduction of consumed energy as the length of allocation slot increase for 3 different length of Random Access Period.

This is to be expected as the device life time depends on the duration of allocation slot and also on the number of slot in a beacon. (as demonstrated in the equation 12).

2) *Energy consumption when scheduled access mode is used*

Fig.8(a) and Fig.8(b) below, show the energy consumption achieved for variable possible scheduled allocation slots and random access slots.

For the first case when the number of slot in a beacon period is equal to 32, the maximum number of scheduled slot can a node get is 3. The remaining 2 slot will be used as random access slots.

For the second case when the number of slot in a beacon period is equal to 128, the maximum number of scheduled slot can a node get is 12. The remaining 8 slot will be used as random access slots.

From the two graphs below, it can be observed that as the number of scheduled access slot is used as we reduce the energy consumption of all nodes. The high performance (less energy consumption) is obtained when scheduled access length is equal to 3 for the first case and 12 for the second one.

We could see that we have great energy efficiency for both cases using scheduled allocations, compared to the first simulation when only random access and improvised access modes are used. This is because polling access mechanism and scheduled allocation does not let the radios of nodes on all the time. When the polling or scheduled intervals arrive, the node is allowed to sleep after transmitting its packets and waking up when its turn to transmit packets arrives.

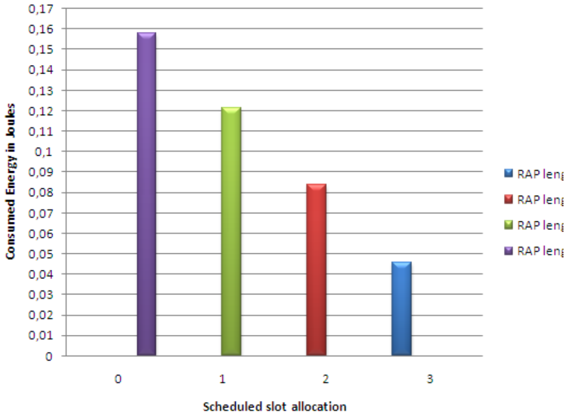


Fig. 8 (a) Average Energy Consumption for different possible length of scheduled access slot and random access Period (beacon period = 32 slot)

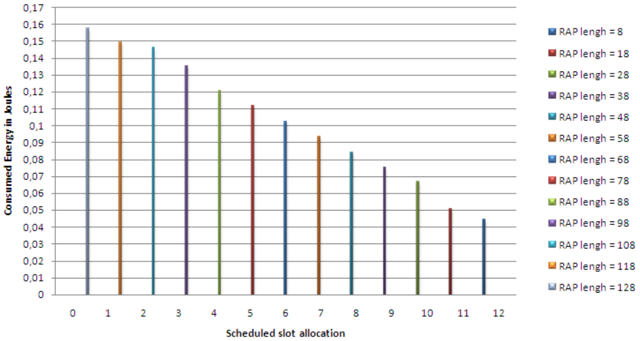


Fig. 8 (b) Average Energy Consumption for different possible length of scheduled access slot and random access Period (beacon period = 128 slot)

VI. CONCLUSION AND FUTURE WORKS

In this paper, we studied MAC access methods used in the IEEE 802.15.6 standard. We evaluated the performances of the different access modes in terms of energy consumption. The results demonstrated that 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] K.S.Deepak. and A.V.Babu, "Optimal Packet Size for Energy Efficient WBAN Under m-periodic Scheduled Access Mode", Proc of IEEE NCC, 2014, pp: 1-6.

[15] Christos Tachtatzis, Fabio Di Franco, David C. Tracey, Nick F. Timmons, Jim Morrison. "An Energy Analysis of IEEE 802.15.6 Scheduled Access Modes for Medical Applications" . ADHOCNETS 2011: 209-222

[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] Tachtatzis C., Di Franco, F., Tracey, D.,Timmons, N.F., Morrison, J.: An energy analysis of the IEEE 802.15.6 scheduled access modes. IEEE GlobecomWorkshop on Mobile Computing and Emerging Communication Networks (MCECN), December, (2010)

[19] http://castalia.npc.nicta.com.au

[20] 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.

[21] 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.

[22] SMA-WiBAN, "MAC and Security Baseline Proposal", IEEE 802.15 Documents, Document no. 196, rev.2, Mar 17th 2010, https://mentor.ieee.org/802.15/documents?is_group=0006.

Designing a jitter buffer for QoS improvement in VoIP networks

Negar Chehregosha
Dept. of Electrical and Computer Engineering,
Science and Research Branch, Islamic Azad
University, Tehran, Iran

Mohammad Ali Pourmina
Dept. of Electrical and Computer Engineering,
Science and Research Branch, Islamic Azad
University, Tehran, Iran

Abstract---Today main challenge in IP networks engineering is simultaneous support of different applications such as sending voice, video and data, with appropriate quality of service. The generated traffic by IP telephone, voice and video conference and on line applications, are real time and time sensitive. Jitter is an usual problem in quality of service of VoIP network. The purpose of this paper is to reduce jitter to improve quality of service.

Achieve Real time voice quality is required jitter smoothing in receiver that usually is done by jitter buffer mechanism. Here we introduce an algorithm to design jitter buffer. We simulate one VoIP network by OPNeT simulator and Matlab software is used to implement the algorithm; then we compare simulation results before and after applying the algorithm and the effects of changes in buffer size on delay and jitter are checked.

Output voice quality will be measured based on PESQ, according to ITU-T P.862 recommendation. The results show packet buffering reduces packets delay and makes values of them become closer together.

Key words VoIP, Jitter, Jitter buffer, Delay, Quality of Service

I. Introduction

SIP is an application layer signaling protocol for creating, modifying and terminating sessions contained with one or more participants. This protocol has been standardized by the IETF and is the most powerful video and voice streaming protocol over internet. SIP uses the elements with the name of “Proxy server” to route requests to current status of users. User registration is also done by SIP and proxy is aware of user’s current location.[1]

Real time transmission consists of two main protocols: RTP and RTCP. RTP is one of the basic principles of VoIP. RTP determines the standard format for packet. This protocol works with RTCP, so that RTCP is used for evaluating transmission characteristics and data streams synchronization. RTP provides arrangements to compensate jitter and out of order data. In summary it can be said, in VoIP technology the voice is transmitted over RTP protocol and control messages are transmitted over SIP protocol, jitter makes sense over RTP.

Delay and jitter are of the most important problems in the network which if go beyond a specified amount, detection will be difficult. One way delay is consisted of the four parameters: propagation delay, transport delay, packetization delay and jitter buffer delay. This is important to design the VoIP network so that don’t tolerate more than 150ms delay. [2] Jitter is variable packets delay to reach destination. This phenomenon has seriously affected the voice quality. Queening delay, packet size and middle links can cause jitter. Jitter will cause packets to arrive out of order, in this case packets may be useless for receiver and they may be discarded, as a result packet will be wasted. Packet loss rate should be less than 5%. Packet loss consists of two parts, first, packets that are lost on the way and never reach their destination. The second set of packets has a great delay; these clusters are discarded by the receiver. Common solution to this problem is using a jitter buffer at the receiver to store the packets for a limited time and sort them.

Buffer overflow occurs if the buffer length is low. The messages are Sent again and traffic increases. If the length of the buffer is high, underflow occurs, link delays and traffic increases on other links. Jitter buffer algorithms are divided into two categories: [3] fixed buffer and adaptive buffer; in fixed buffer a fixed size for jitter buffer is considered. When the network condition is stable, fixed buffer is more effective. Size of the adaptive buffer changes with respect to the measured delay and jitter this means that test packets are sent and received at alternative times and the network parameters can be achieved then buffer size is determined according to the parameters. Better balance between latency and packet lost is advantage of this kind of buffer but the disadvantage is that it causes delay.

In this paper we propose an algorithm to reduce jitter and compensate its effects. Effects of changing size of jitter buffer on delay, jitter and quality of service are examined. In this algorithm, we buffer packets in descending order according to their SN and then send them to the codec. OPNeT software is used for network simulation and algorithm is applied Using MATLAB. The results show applying the algorithm is effective and quality of service increases. In Section 2, we introduce the proposed jitter buffer algorithm. In Part 3 we perform the simulation and offer results for different buffer lengths. Finally, conclusions based on the simulations are described in section 4.

II. The proposed jitter buffer design algorithm

As you know in IP networks, [4], [5], [6], [7], each packet has a sequence number. The idea of our algorithm is packets are received by the receiver are arranged according to their SN. We are trying to arrange packets in decreasing order. It means the packet with a smaller SN placed closer to the beginning of the buffer on entering the codec. N is the length of the buffer, and SN_L is the last packet number entered into the codec. Figure 1 shows the buffer.

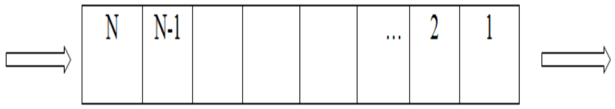


Figure1. How packet enter and exit buffer

The SN which is located in block number 1 is the smallest. When the first packet arrives empty buffer, it will be placed in the buffer farthest block, the block number N. The SN of last packet, which has entered codec is saved, SN of every packet that comes, is compared to the this stored number, if the number of newly received packet is less than the last number entered Codec, Pack will be removed and will not enter into codec. If it was smaller than SN_L+M, will enter the buffer but if it was greater than SN_L+M, will be removed. It means it must be $SN_L<SN\le SN_L+M$. The value of M is proportional to the length of the buffer.

To sort the packets selection sort method is used. Sorting operation takes place within its list without need for auxiliary memory. This algorithm sorts the list in ascending or descending order within a few steps. This means that at each stage, the largest (or smallest) element is found and moved to the bottom of the list. When the buffer is 70 percent full packets sorted. When entering a new packet, it has to be checked buffer is empty or not? To do this, we define a parameter read & write. When one packet added to buffer this parameter is increased and when a packet enters the codec it is decreased. Its initial value is zero. When entering a packet this pointer will be checked if it is less than N there is empty place in buffer and new packet can be entered. If the value of the pointer is N, packet will be eliminated. In other words the buffer overflowed. Flowchart of algorithm is presented in figure 2.

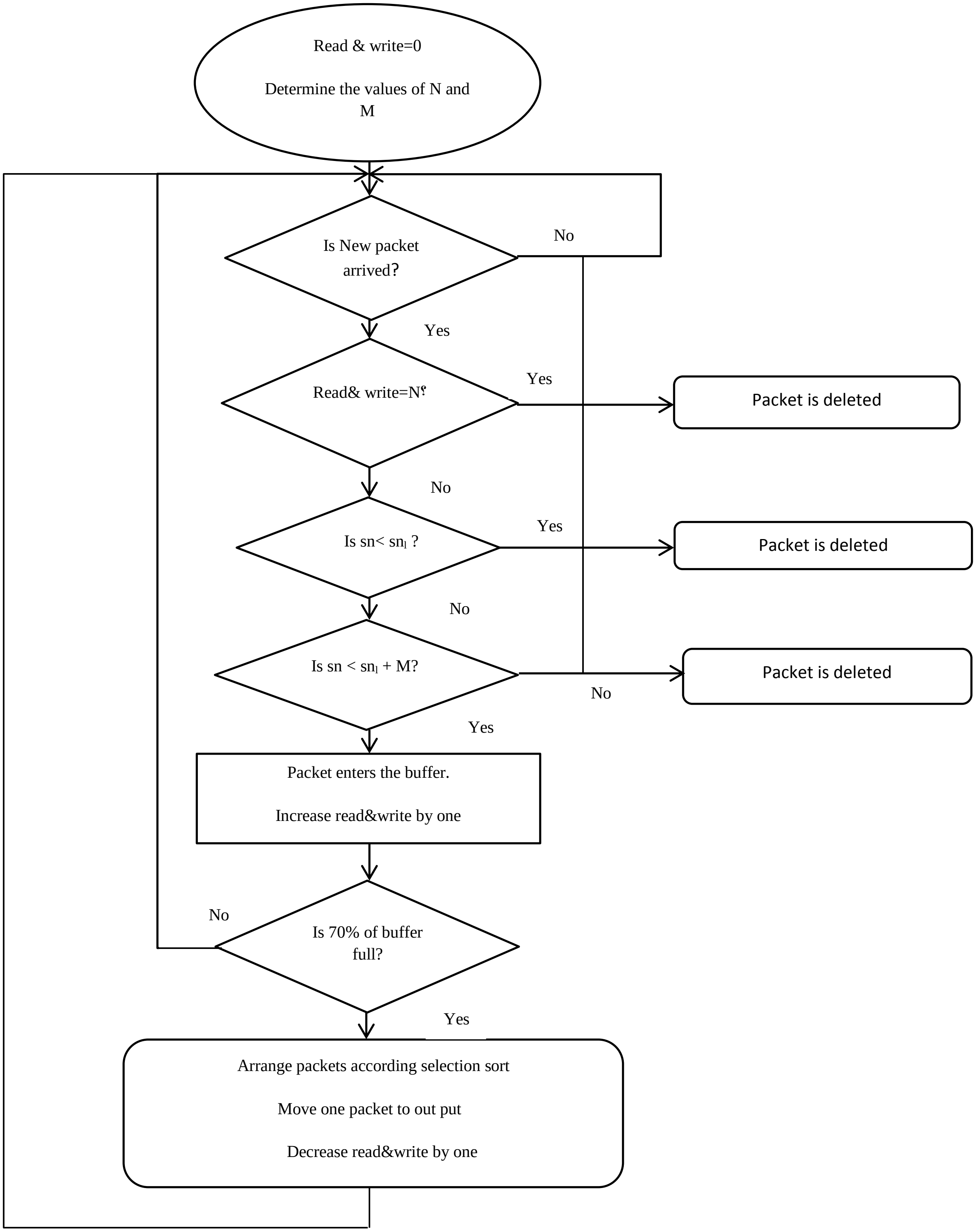


Figure2. Flowchart of the proposed algorithm

III. Simulation and Results

There are several items affect performance of network and Quality of Service that can be considered in the simulation such as packet length, background traffic, jitter buffer and codec. Here mentioned parameters are considered fixed except jitter buffer and only effects of jitter buffer algorithm will be checked. New algorithm is examined to evaluate delay, jitter, and service quality. These parameters for the non-buffer and buffer length of 4, 7 and 10 are simulated and compared.

A. Simulation settings

The OPNeT v14.5 simulator and Matlab R2008b are used for simulation. As we said, VoIP traffic generated by OPNeT is given as input to Matlab and using MATLAB software, algorithm is applied. Network scenario is presented in figure 3.

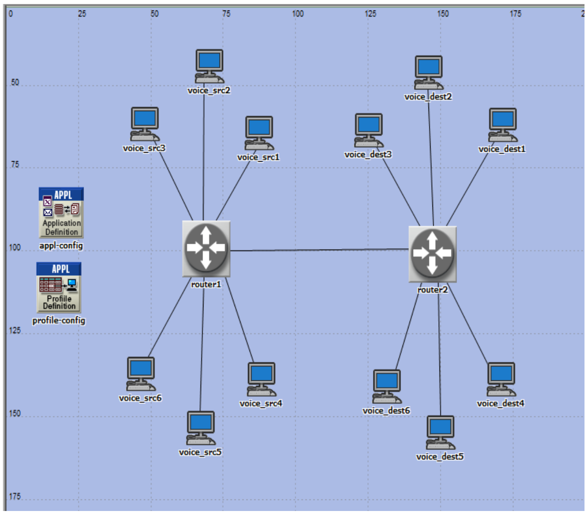


Figure 3. Network scenario

In the network scenario SIP signaling protocol, G.723.1 codec are used and the data transmission rate is 2Mbps. Nodes are fixed. And simulation was done in 30 minutes or 1800 MS, delay and jitter were measured for 100 samples, every 18 MS.

B. Simulation results for networks with different buffer lengths

Figure 4 shows end-to-end delay for buffer with lengths of 0, 4, 7 and 10 and Figure 5 indicates jitter for them. In all plots at intervals when the chart is empty, there is no traffic exchanged. As you can see from the delay curves due to the codec, delay is at least 40 milliseconds.

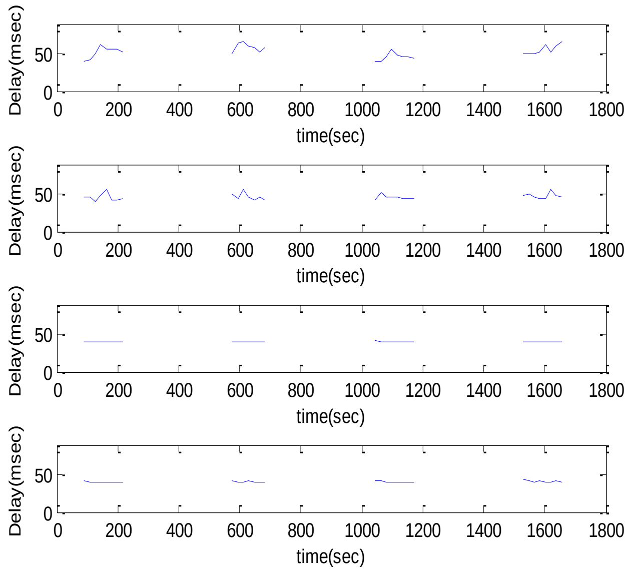


Figure 4. end-to-end delay for buffer with lengths of 0, 4, 7 and 10 for 30 minute

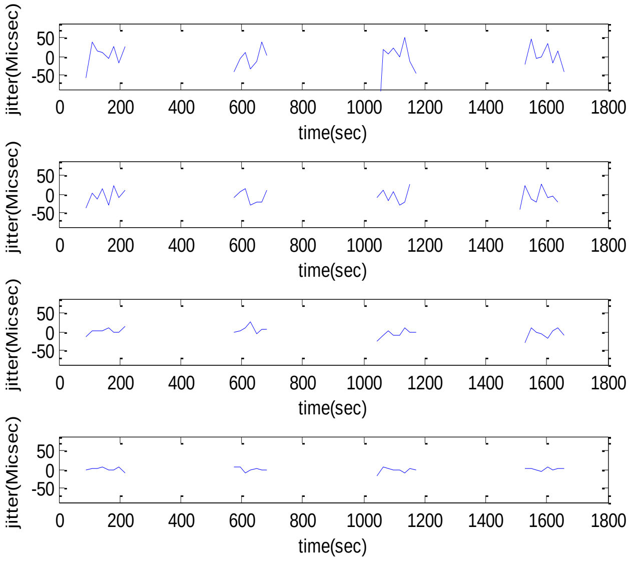


Figure 5. jitter for buffer with lengths of 0, 4, 7 and 10 for 30 minute

The graphs clearly indicate that by increasing the buffer size delay decreases and its graphs became smoother. Smoother delay curves mean delay values become close together. In other words, jitter is reduced. The jitter diagram also indicates low jitter for larger buffer size. In the case of N = 0 packets arrive destination without buffering but in three next cases packets are arranged N to N. So it is seen that the results are improved. With larger buffer it will be possible to access late packets and prevent losing them. As you know, jitter and delay are the most important factors of service quality. As a result with reduction of these factors the quality of service will increase in the network. Table 1 shows the average values of delay, the mean absolute jitter values and MOS obtained from different length of the buffer and you can see jitter curve in Figure 6, delay curve in Figure 7 and MOS curve in Figure 8.

N	Jitter (micro sec)	Delay (milli sec)	MOS
0	34	54	3.07
4	24	49	3.29
7	10	45	3.55
10	4	42	3.68

Table1. Values of Jitter, Delay and MOS for different length of the buffer

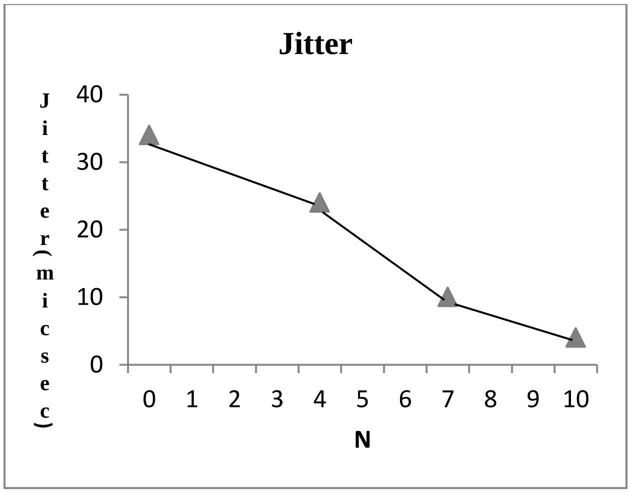


Figure 6. Mean absolute jitter values for different length of the buffer

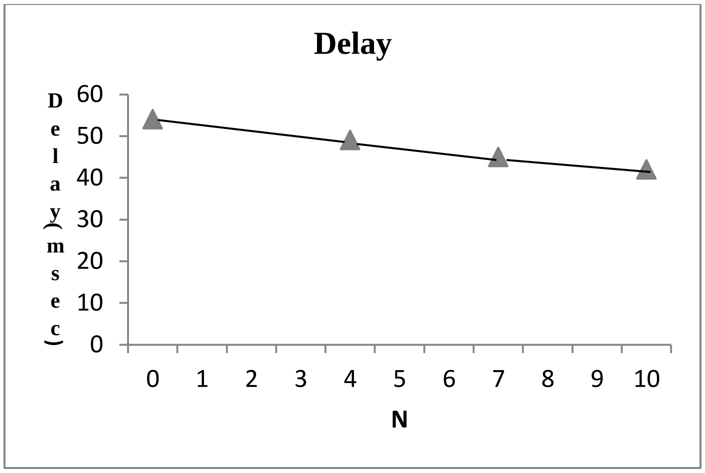


Figure 7. Average values of delay for different length of the buffer

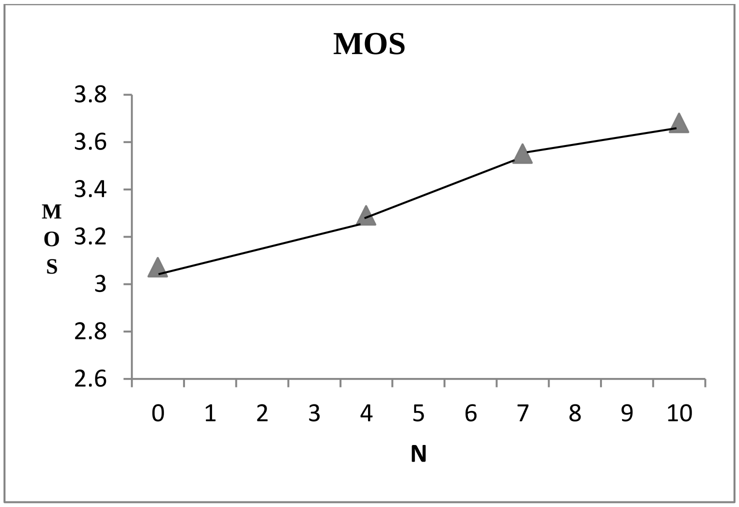


Figure 8. MOS values for different length of the buffer

IV. Conclusions and recommendations

Buffer size was changed from 0 to 15, graphs were plotted for values of N that their changes were manifest. Results for the buffer size larger than 10 was almost identical. According to the graphs, it seems optimal buffer size for this network is 10. According to the results of simulations, because the algorithm sorts packets N by N, it prevents resending the out of order packets and late packet, so it is effective in minimizing traffic, delay and jitter.

As we know it takes time to buffer packets so the proposed algorithm somewhat increases delay but as is clear from the results its beneficial effect is more impressive than its bad effect.

Proposed jitter buffer was fixed and non-adaptive; buffer size was given algorithm manually. Therefore the design and implementation of adaptive jitter buffer is recommended, the buffer with variable

References

[1] Maryam Kiani, “Analysis of Real Time Communication in Next Generation Network with Regard to Quality of service”, Dept. of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran, Thesis for receiving «M.Sc» degree on electronics engineering , 2010.

[2] S. Sahabudin, and M.Y. Alias , “End-to End Delay Performance Analysis of Various Codecs on VoIP Quality of Service”, IEEE 9th Malaysia International Conference on Communications (MICC), pp: 607 – 612 , 2009.

[3] Chunxia Tu , “Study on QoS Protection Mechanism of VoIP Systems”, 2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC), Publisher: IEEE, Hubei, pp: 151 – 153, 2011.

[4] M. Broitman, N. Shilinskii and, K. Solovyov, “ Adaptive Management Algorithms for a Fixed Jitter Buffer”, Automatic Control and Computer Sciences, Vol. 46 No. 1, pp. 12–17, 2012.

[5] M. Baratvand , M. Tabandeh, A. Behboodi, and A. Fotowat Ahmadi, “Jitter-Buffer Management for VoIP over Wireless LAN in a Limited Resource Device” . Fourth International IEEE Conference on Networking and Services, Gosier, pp: 90 – 95, 2008.

[6] S. Paulsen , T. Uhl and K. Nowicki, “Influence of the Jitter Buffer on the Quality of Service VoIP”, 3rd International IEEE Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT), Budapest, pp: 1-5, 2011.

[7] O. Slavata, Holub, J. and P. Hübner, “Impact of Jitter and Jitter Buffer on the Final Quality of the Transferred Voice” IEEE 1st International Symposium on Wireless Systems (IDAACS-SWS), Offenburg ,pp: 120 – 123, 2012.

AUTHORS PROFILE

Negar Chehregosha was born in Kermanshah, IRAN in 1986. She received the B.S. degree in Electronics Engineering from Razi University, Kermanshah, IRAN in 2009. She is a M.S. student in Communications Engineering at science and Research Branch, Islamic Azad University, Tehran, Iran, now. Her major fields of interest are networks communications and VoIP. She is currently working as a VoIP expert.

Mohammad Ali Pourmina received his B.S and M.S. degrees in electrical engineering (Telecommunication) from Islamic Azad University, Central branch, Tehran, Iran, in 1989 and 1991respectively.He received his Ph.D. degree at Science and research Branch, Islamic Azad University, Tehran, Iran, in 1996. He joined Science and research Branch, Islamic Azad University, Tehran, Iran in 1996. He has been a member of Iran Telecommunication research center since 1992. He has performed research in the areas of packet radio networks and digital signal processing systems since 1991. His current research interests include spread-spectrum systems, cellular mobile communications, indoor wireless communications, DSP processors, Cross layer design and wireless multimedia networks.

Trend Analysis in Academic Journals in Computer Science Using Text Mining

* Adebola K. Ojo and Adesesan B. Adeyemo
[1,2]Department of Computer Science
University of Ibadan
Ibadan, Nigeria
[*]Corresponding author

***Abstract*— Text mining is the process of discovering new, hidden information from texts- structured, semi-structured and unstructured. There are so many benefits, valuable insights, discoveries and useful information that can be derived from unstructured or semi- unstructured data. In this study, text mining techniques were used to identify trends of different topics that exist in the text and how they change over time. Keywords were crawled from the abstracts in Journal of Computer Science and Technology (JCST), one of the ISI indexed journals in the field of Computer Science from 1993 to 2013. Results of our analysis clearly showed a varying trend in the representation of various subfields in a Computer Science journal from decade to decade. It was discovered that the research direction was changing from pure mathematical foundations, Theory of Computation to Applied Computing, Artificial Intelligence in form of Robotics and Embedded Systems.**

Keywords-component; Computer Science, Text Mining, mathematical foundations, applied computing, Robotics, Embedded Systems

I. Introduction

Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources [1]. *Previously unknown* implies discovering genuinely new information. *Unstructured* means free naturally occurring texts as opposed to HyperText Markup Language (HTML), eXtensible Markup Language (XML), and other scripting languages. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down [2]. [3] referred to it as a collection of methods used to find patterns and create intelligence from unstructured data.

Text mining techniques are used to draw out the occurrences and instances of key terms in large blocks of text, such as articles, web pages, complaint forums, or Internet chat rooms and identify relationships among the attributes [4]. Often used as a preparatory step for data mining, text mining often translates unstructured text into a useable database-like format suitable for data mining for further and deeper analysis [5]. [3] also described text mining as an emerging technology that can be used `to augment existing data in corporate databases by making unstructured text data available for analysis.

Generally research publications have been on the increase globally. As a result, different areas are being covered such as science, engineering, agriculture, medicine and education. However, it is a time consuming task to determine manually the areas being focused on by authors who are publishing in these journals. In this study, a framework for discovering the research trends in Computer Science in the last three decades was developed using text mining techniques.

This work used an ISI indexed journal called Journal of Computer Science and Technology (JCST). The trending of topics published in papers in JCST across two decades was explored.

There are three types of textual data for text mining. These include Title of the paper, Abstract of the paper and complete body of the paper [6]. The data used in this study were the abstracts with the keywords. Analysing the abstract of a paper is appropriate since it contains the detailed objective of a paper and did not contain extraneous items such as tables and images [6]. Three data sets were created with the number of observations (that is, paper abstracts) as shown in Figure 1.

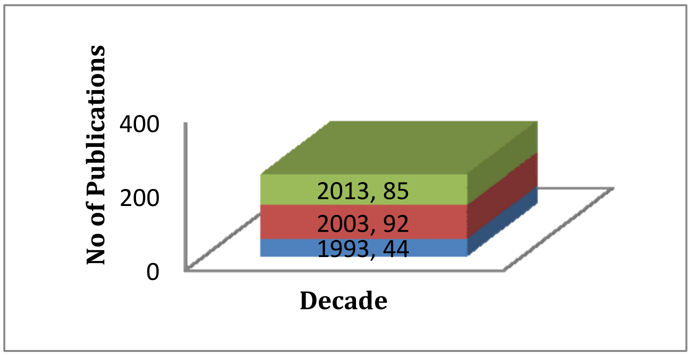


Figure 1: Number of Paper Abstracts in each Period

II. Materials and Method

The framework for Research Trend discovery is presented in Figure 2.

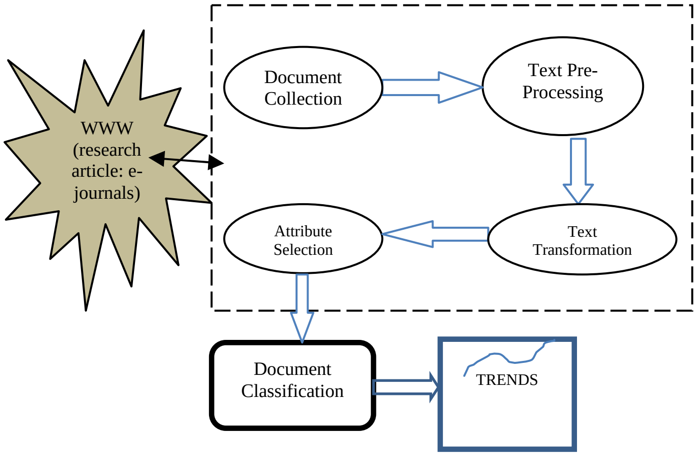


Figure 2: Framework for Academic Journal Articles (Trend Discovery)

Document Collection: The first phase in Figure 2 is the document collection. In this phase, the abstracts of the academic journal (JSCT) were *extracted* as documents (doc, pdf and html) crawled from the internet. These text documents were stored in different formats (pdf, doc, txt, html and xls). This depends on the nature and type of the text data. These contained the list of the abstracts of the volumes published in the years 2013, 2003 and 1993.

Text Pre-processing: This is also known as tokenization or text normalization. It involves the process of text clean-up (advertisements from the web pages are removed as well as the tables, figure and formulas) and tokenization (splitting up a string of characters into a set of tokens). During term extraction, character text was first parsed into words. This process also stripped away words that conveyed no meaning. Adjectives, adverbs, nouns and multi-word were extracted from the document. Noisy data, such as, tags, punctuation marks, white spaces, special characters and digits were extracted as well. Also, certain words occurred very frequently in text data. Examples included “the” and “a”. These words were removed from the term collection because they had no meaningful content.

Text Transformation/Attribute Selection: By creating a list of stop words and eliminating them, the number of indicator variables created was reduced. After removal of stop words, stemming was performed. Word frequency and inverse document frequency were two parameters used in filtering terms. Low term frequency (TF) and document frequency (DF) terms were removed from the indexing of those documents. In “Bags of words” representation each word is represented as a separate variable having numeric weight.

III. Results and Discussion

The abstracts were extracted from http://link.springer.com/journal/volumesAndIssues/11390. One well known subject classification system for Computer Science is the ACM Computing Classification System devised by the Association for Computing Machinery [7] [8]. Computer Science was divided into ten (10) subfields. The subfields included Algorithm and Data Structures, Artificial Intelligence, Communication and Security, Computer Architecture, Computer Graphics, Databases, Programming Languages and Compilers, Scientific Computing, Software Engineering, and Theory of Computation. Table 1 presents the article classifications for 1993, 2003 and 2013 while Table 2 shows percentage distribution of article classifications.

In the Text Extraction Process, all the abstracts were parsed into independent words. This process also stripped away words that conveyed no meaning. Adjectives, adverbs, nouns and multi-word are extracted from the document. Noisy data, such as, tags, punctuation marks, white spaces, special characters and digits were extracted as well. Table 1 shows the clusters generated with the term frequencies and weights. Our suggested cluster labels were based on the descriptive terms and corresponding fields.

TABLE 1: ARTICLE CLASSIFICATIONS FROM 1993 TO 2013

No	Descriptive Terms	1993	2003	2013
1	Algorithms, Data Structures	33	27	24
2	Artificial Intelligence, Automated Reasoning, Computer Vision, Natural Language Processing, Machine Learning, Robotics	13	37	112
3	Networking, Computer Security, Cryptography, Concurrent, Parallel & Distributed Systems	51	128	221
4	Computer Architecture, Operating Systems	17	70	157
5	Computer Graphics, Image Processing	19	49	67
6	Relational Databases, Data Mining	12	41	41
7	Compiler Theory, Programming Language Pragmatics, Programming Language Theory, Formal Semantics	49	6	20
8	Computational Science, Numerical Analysis, Symbolic Computation, Computational Chemistry, Bioinformatics & Computational Biology, Computational Neuroscience	1	1	10
9	Software Engineering, Formal Methods, Algorithm Design, Computer Programming, Human-Computer Interaction, Reverse Engineering	0	5	34
10	Theory of Computation, Automata Theory, Computability Theory	33	50	91

TABLE 2: PERCENTAGE DISTRIBUTION OF ARTICLE CLASSIFICATIONS FROM 1993 – 2013

No	Descriptive Terms	1993	2003	2013
1	Algorithm and Data Structures	14.5	6.5	3.1
2	Artificial Intelligence	5.7	8.9	14.4
3	Communication and Security	22.4	30.9	28.4
4	Computer Architecture	7.5	16.9	20.2
5	Computer Graphics	8.3	11.8	8.6
6	Databases	5.3	9.9	5.3
7	Programming Languages and Compilers	21.5	1.4	2.6
8	Scientific Computing	0.4	0.2	1.3
9	Software Engineering	0.0	1.2	4.4
10	Theory of Computation	14.5	12.1	11.7
	TOTAL (%)	**100**	**100**	**100**

A. Trends

Trend analysis is used for identifying trends in documents collected over a period of time [2]. Identification of meaningful patterns and trends and the extraction of potential knowledge in large volumes of text data is an important task in various fields [9][10]. The appearances of specific terms across the two decades are used to understand the trends and research patterns of sub fields in Computer Science. A frequency value of 'n' for a term means that particular term was mentioned in the abstracts of 'n' distinct journals. Figure 3 shows the percentage of papers contributed for the ten disciplines across the period.

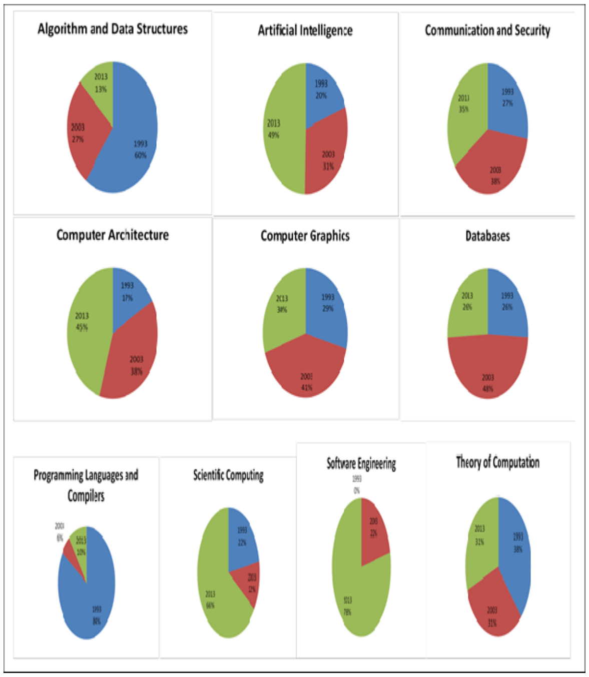


Figure 3: Percentage of Papers Contributed For Ten Computer Science Disciplines From 1993 To 2013

In Figures 6, most of the papers in 1993s (22.4%) were presented on Communications and Security followed by Programming Languages and Compilers (21.5%), Theory of Computation and Algorithms and Data Structures (14.5%), Computer Graphics (8.3%), Computer Architecture (7.5%), Artificial Intelligence (5.7%), Databases (5.3%), Scientific Computing (0.4) and Software Engineering (0.0%). The same trend was not observed in the following years. Percentage of papers published in Artificial Intelligence, Communication and Security, Computer Architecture, Computer Graphics, Databases and Software Engineering gradually increased from 1993 through 2003 (as shown in Figure 5) while publications in Algorithm and Data Structures, Programming Languages and Compilers, Scientific Computing and Theory of Computation reduced drastically (as shown in Figure 6). However In 2013, there was a great increase in papers published in Artificial Intelligence, Computer Architecture and Software Engineering while in there was relatively fewer numbers of papers published in the other disciplines. This shows that the research direction is changing from pure mathematical foundations, Theory of Computation to applied computing, Artificial Intelligence in form of Robotics and embedded systems.

Figure 4 shows the trend plot of the sub fields in Computer Science across the two decades.

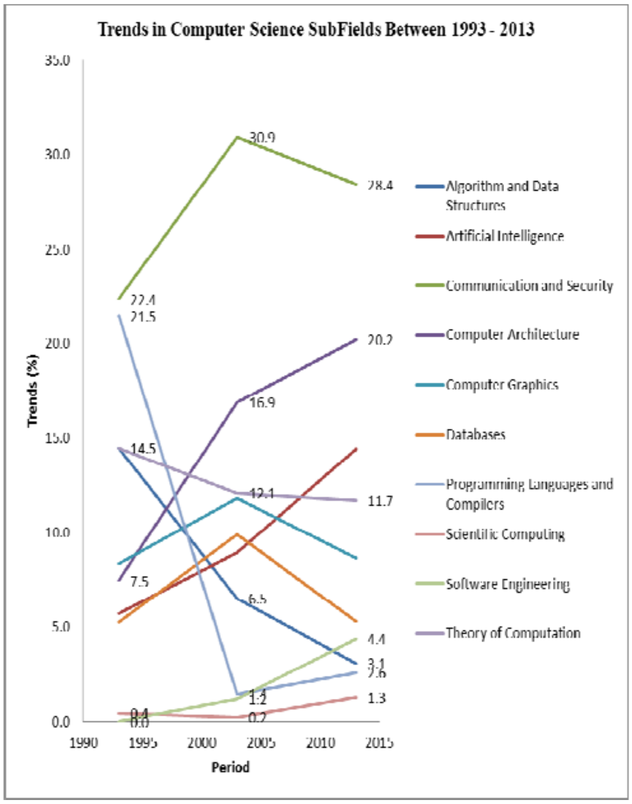


Figure 4: Trend Plot of the Sub Fields in Computer Science across the two decades

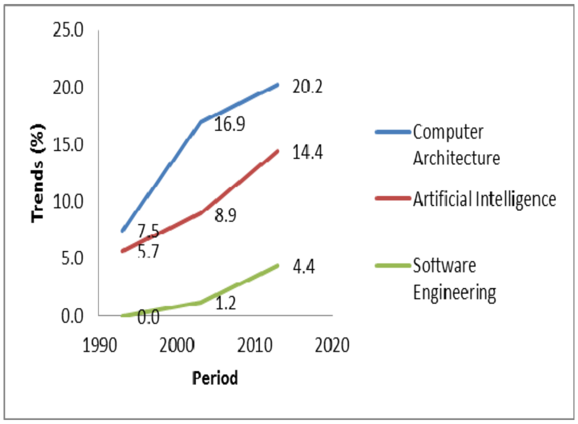


Figure 5: Gradual increase of some sub fields across the two decades

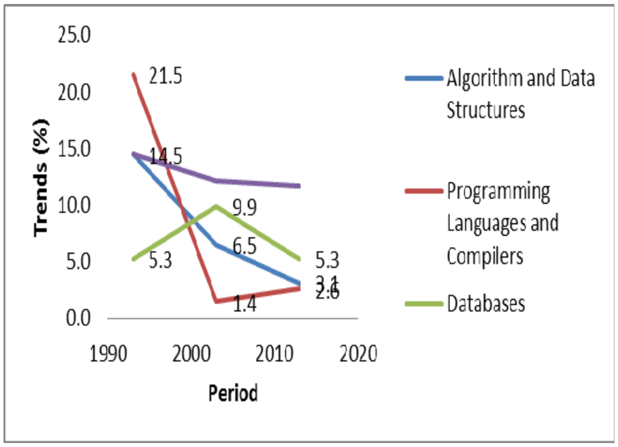


Figure 6: Gradual decrease of some sub fields across the two decades

Figure 6 shows the trend plot of various sub fields in Computer Science with much less representation in the papers compared to the large scale representation of the discipline as shown in Figure 5.

TABLE 3: PERCENTAGE INCREASE OVER THE TWO DECADES

Descriptive Terms	% Increase over Decades
Algorithm and Data Structures	21.34
Artificial Intelligence	252.81
Communication and Security	127.16
Computer Architecture	271.00
Computer Graphics	103.47
Databases	100.26
Programming Languages and Compilers	11.98
Scientific Computing	293.44
Software Engineering	>300.44
Theory of Computation	80.92

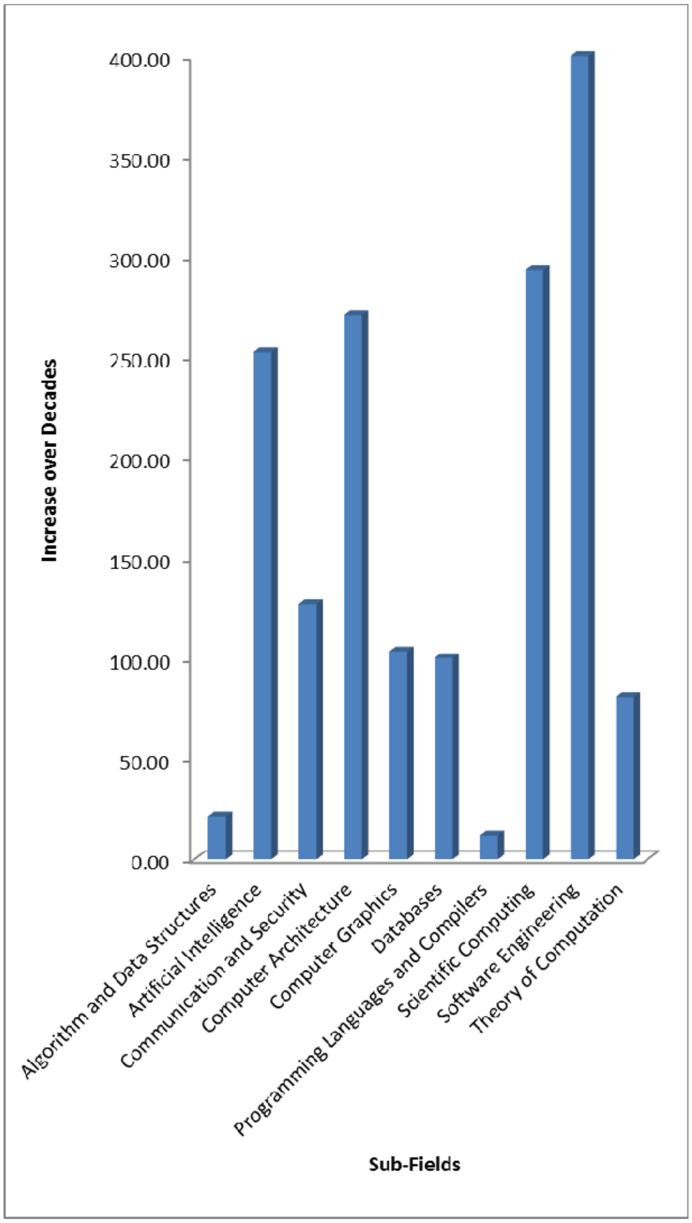


Figure 7: Percentage Increase over the Two Decades

Table 3 and Figure 7 showed percentage increase over the two decades of various disciplines in JCST. Computer architecture, artificial intelligence, scientific computing, software engineering and communication and securities were the disciplines where highest percentage increase was recorded. The least were databases and computer graphics. It is widely known that the growth of computer hardware: processors, embedded systems (such as, mobile devices) and controllers occurred in 2000s and hence we can expect more papers published in Computer Architecture and Artificial Intelligence during the decade of 2003. It is gratifying to observe that trend in the plot (Figure 7).

IV. CONCLUSION

In this work, text mining is applied to figure out trends in research topics related to various subfields in Computer Science academic journal articles within the period of two decades. This analysis can also be extended to find trends in research topics related to other disciplines in the academic journal articles. A similar approach can also be used to analyse many academic electronic journal articles (corpus) in other fields. Text mining has tremendous potential in identifying trending topics during a period of time.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of [1] M. Hearst. *Untangling Text Data Mining,* in the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[2] V. Gupta and G. S. Lehal, *A Survey of Text Mining Techniques and Applications*, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No.1. August 2009.

[3] F. Louise, and M. Flynn, *Text Mining Handbook*, Casualty Actuarial Society E-Forum, 2010.

[4] D. Robb , *Taming Text,* Retrieved from http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e8bd354a88a41ad89a928d23360302a4959035699f17e2ba8a63e2dd032c73f8a7fmt=H, 2005

[5] P. Cerrito, *Inside Text Mining,* Retrieved from http://wilsontxt.hwwilson.com/pdffull/06619/275n6/gs9.pdf, 2005.

[6] Z. Shaik, S. Garia, and G. Chakraborty, *SAS® Since 1976: An Application of Text Mining to Reveal Trends*, Proceedings of the SAS Global Forum 2012 Conference, SAS Institute Inc., Cary. support.**sas**.com/resources/papers/proceedings12/135-2012.pdf, 2012

[7] B. Mirkin, S. Nascimento, L. M. Pereira, *Representing a Computer Science Research Organization on the ACM Computing Classification System*, in Eklund, Peter; Haemmerlé, Ollivier, Supplementary Proceedings of the 16th International Conference on Conceptual Structures (ICCS-2008), CEUR Workshop Proceedings**354**, RWTH Aachen University, pp. 57–65, 2008.

[8] Wikipedia: http://en.wikipedia.org/wiki/Outline_of_computer_science

[9] A. Kao and S. R. Poteet, (Eds), *Natural Language Processing and Text Mining*, Springer, London, UK, 2007.

[10] S. G. Cho, and S. B. Kim, *Identification of Research Patterns and trends Through Text Mining*, International Journal of Information and Educational Technology, 2(3). June 2012.

AUTHORS PROFILE

Adebola K. OJO is a PhD student in the Department of Computer Science, University of Ibadan, Nigeria. She is a registered member of the Computer Professional of Nigeria (CPN) and Nigeria Computer Society (NCS). She had her BSc in Computer Engineering from Obafemi Awolowo University, Nigeria. She also obtained her Masters of Science Degree in Computer Science from University of Ibadan, Nigeria. Her research interests are in Digital Computer Networks, Data Mining, Text Mining and Computer Simulation. She is also into data warehouse architecture, design and data quality via data mining approach.

Dr. Adesesan Barnabas Adeyemo is a Senior Lecturer at the Computer Science Department of the University of Ibadan, Nigeria. He obtained his PhD, M.Tech., and PGD Computer Science degrees at the Federal University of Technology, Akure. His research activities are in Data Mining, Data Warehousing & Computer Networking. He is a member of the Nigerian Computer Society and the Computer Professionals Registration Council of Nigeria. Dr. Adeyemo is a Computer Systems and Network Administration specialist with expertise in Data Analysis and Data Management.

d-HMAC — An improved HMAC algorithm

Mohannad Najjar
University of Tabuk
Tabuk, Saudi Arabia
najjar@ut.edu.sa

Abstract—**The keyed-hash message authentication code (HMAC) algorithm is a security tool primarily used to ensure authentication and data integrity in information systems and computer networks. HMAC is a very simple algorithm, and relies on hash functions that use a secret key. HMAC's cryptographic strength is based on the use of effective cryptographic characteristics such as balancing and the avalanche effect. In this study, we develop a new algorithm, entitled dynamic HMAC (d-HMAC), to improve and enhance the cryptographic characteristics of HMAC. The improved algorithm provides stronger resistance against birthday attacks and brute force attacks. To achieve this objective, HMAC constant values *ipad* and *opad* are dynamically calculated in d-HMAC. Values for *ipad* and *opad* will be obtained from the HMAC input message, the public key of the receiver, and a substitution-box (*S*-box) table with enhanced security characteristics specifically created for this purpose. We demonstrate that the improved d-HMAC algorithm is more resistant to known cryptographic attacks, and prove that it exhibits similar or better cryptographic characteristics than HMAC.**

Keywords-cryptography; data integrity; authentication; MAC; HMAC; hash functions; SHA-256

I. Introduction

Authentication is one of the primary aspects of security, because it confirms the identity of the source and ensures that data has not been altered. Message authentication codes (MACs) are one of the most important authentication and data integrity tools. In this study, we will improve the functionality of HMAC, which is a type of MAC that uses hash functions.

Compared with other MAC types, HMAC is considered more effective, as described by [12]:

- MAC uses encryption algorithms that are relatively slow.
- Many hardware cryptographic tools are built to manage large volumes of data.
- Many cryptographic algorithms require licenses, whereas HMAC is free of charge.

.

HMAC is a cryptographic tool that ensures authentication and data integrity [2,5]. HMAC is used in data exchanging and warehousing, to ensure the validity of the source. HMAC is a specific type of MAC function that can use any type of hash function that uses a secret key shared between two parties, to process an input message *m*. HMAC is one of the most prominent cryptographic algorithms, and is used to ensure that saved or exchanged data is not changed (intentionally or accidently) without authorization. HMAC uses two values, *ipad* and *opad*, to ensure that keys are pseudorandom. These values are fixed and known, which causes security weaknesses against known cryptographic attacks [13].

In this study, we present an improved d-HMAC algorithm, which uses dynamic *ipad* and *opad* values to provide more robust security than HMAC. Dynamic *ipad* and *opad* values increase the d-HMAC algorithm's resistance to known cryptographic threats such as birthday and brute force attacks. Furthermore, d-HMAC is an effective tool for creating pseudorandom initial values for hash functions.

Cryptographic tests will be conducted against different types of security weaknesses; this will show the effectiveness of d-HMAC compared with HMAC, and demonstrate d-HMAC's useful cryptographic properties, including balance and avalanche effects.

List of important symbols used in this paper (National Institute of Standards and Technology, 2002):

h	Hash function: SHA-256.
b	Number of bits in block *h*.
IV	Hash function initial value.
m	Input data processed by HMAC.
Yi	Message *m i*th blocks, $0 \le i \le (l\text{-}1)$.
l	Quantity of blocks in message *m* after bit padding.
N	Lengths of message digest.
K	Secret key, if *K* length > *b* then $K = h(K)$.
K+	*K* padded with zeros.
ipad	Inner pad of (36_H) reiterated *b*/8.
opad	Outer pad: $(5c_H)$ reiterated *b*/8.
h(m)	Message digest for d-HMAC with n bits length.
Y	All possible message digests.
$\oplus$	XOR operation (bitwise exclusive-OR).

This paper is organized as follows: in Section 2, hash functions are introduced; in Section 3, HMAC functions are explained; in Section 4, the improved d-HMAC algorithm is described; in Section 5, test results are presented; in Section 6, d-HMAC's improved resistance to cryptographic attacks is presented; in Section 7, we present our conclusions.

II. Hash functions

A hash function is a cryptographic tool used to ensure the integrity of data by preventing unauthorized or accidental modifications. A hash function may also be called a manipulation detection code (MDC). Moreover, hash functions are also used for other cryptographic applications, such as digital signatures and password storage [8,11].

A hash function is a function $h: M \rightarrow Y$ that must fulfill the following requirements (where $m \in M, y \in Y$):

- It compresses message m with a distinct length to message digest $h(m) \in Y$ with a fixed length.
- It can straightforwardly compute message digest $h(m)$ for any message m.
- It is computationally infeasible to compute any $m' \in M$ for most $y \in Y$ where $y = h(m')$ (one-way characteristic).
- It is computationally infeasible to compute another message m' for message m where $h(m) = h(m')$ (pre-image resistance characteristic).
- It is infeasible to compute two different messages m and m' where $h(m) = h(m')$ (Second pre-image resistance characteristic).

The hash function transformation "Fig. 1" for message $m = m_1||\mathrm{m}_2||\ldots||m_t$ is divided into fixed length blocks $m_1, m_2, \ldots, m_t$ and can be described as follows:

$$H_0 = IV$$

$$H_i = \varphi(m_i, H_{i-1}), \text{ Where } i = 1,2,...,t$$

$$h(m) = H_t$$

Where message m and fixed initial value IV are the input of the hash function, H_i is the chain variable calculated by compression function φ and the ith block of m, and the output result is $h(m)$. $h(m)$ has different names in the cryptography literature, including hash result and fingerprint [10,11]. In this paper, it is referred to as the message digest. The hash function structure is depicted in Fig. 1 [8].

HMAC and d-HMAC can use any type of hash function, including MD5, RIPEMD-160, SHA-256, SHA-512, and PETRA. Most hash functions have constant initial values, except the PETRA hash function, which has dynamic initial values [3]. In this paper, we will focus on the SHA-256 hash function; thus, we will conduct all of our d-HMAC tests using SHA-256 [6].

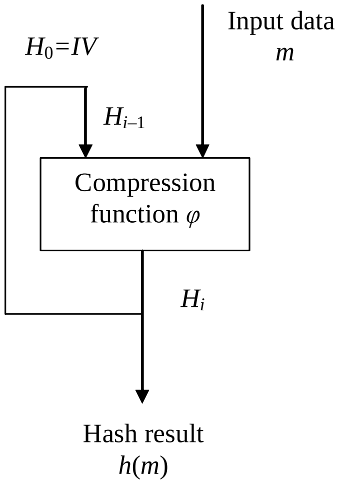


Figure 1. General model of the hash function *h*

III. KEY-HASHED MESSAGE AUTHENTICATION CODE (HMAC)

Authentication is one of the primary aspects of security [17], and ensures the authenticity of senders, receivers, and data. It confirms that senders and receivers are who they claim to be. MAC is one of the main cryptographic tools used to ensure authentication. MAC can be constructed by using block cipher algorithms, or by using a hash function that employs a secret symmetric key. In this work, we will focus on HMAC, which is a specific type of MAC that uses hash functions and secret keys.

The main design objectives of HMAC functions are as follows [2]:

- HMAC can use any type of hash function as it is. Most hash functions are available free of charge.
- HMAC does not interfere with the hash function, thus its performance is not negatively affected.
- HMAC prefers to use readily available keys in a straightforward manner.
- HMAC's algorithm is simple and easy to modify, allowing the security level or speed of the underlying hash function to be upgraded if necessary.

We assume that HMAC can use any hash function h without modifications and secret key K. Hash function h will execute based on compression function φ for a message m containing l blocks. The length of each block l in bits is denoted by b, which indicates that $l * b$ will equal the length of message m after bit padding. We denote by n the length of the message digest in bits. In our research, we use SHA-256 functions, which results in n = 256. The length of shared secret key K can equal b or less; for keys longer than b bits, a hashing computation must be performed using the h function. HMAC's designers recommend using a K that is at least n bits long. Using keys with lengths less than n is discouraged, because it reduces the strength of HMAC's security functions.

HMAC's designers recommend the use of highly random keys, and they recommend changing keys as a standard security practice. This minimizes the negative consequences of exposed keys, and mitigates threats that depend on the collection of data calculated by HMAC using the same key, such as offline brute-force attacks.

It is advantageous to improve hash functions to use dynamic initial values *IV* instead of fixed values. Furthermore, HMAC calculates intermediate values of ($K^+ \oplus opad$) and ($K^+ \oplus ipad$) once for the same *k*. These intermediate values will be used many times to authenticate the same key; as a result, these values, including the secret keys, must be protected against any type of disclosure [15].

The main equation of HMAC is defined as follows, Fig. 2:

$$HMAC_K(m) = h\left(K^+ \oplus Opad, h\left(K^+ \oplus Ipad, m\right)\right) \quad (1)$$

HMAC can be calculated using the following steps [2]:

1. K^+ is calculated by padding zeros onto *K*'s left side to increase its length to *b*-bits.
2. Calculate S_i of *b*-bit length by XORing *ipad* with K^+.
3. Concatenate S_i with *m* to be equal to *m'*.
4. Calculate message digest *h*(m') for *m'* by using *h*.
5. Calculate S_0 of *b*-bit length by XORing *opad* with K^+.
6. Concatenate *h*(*m'*) with S_0 to be equal *m"*.
7. Calculate message digest *h*(*m"*) for *m"* by using *h* to get the final result for HMAC.

We can conclude that the main objective of using distinct constant values for *ipad* and *opad* in HMAC while calculating the message digest twice is to avoid cases in which:

$K^+ \oplus ipad$ OR
$K^+ \oplus opad$ will be a string of zero values

HMAC algorithm keys must have cryptographic characteristics that define their length, randomness, and complexity [2]. In terms of HMAC design principles, the bit length of key *K* can be any arbitrary size. The optimal size of *K* is equal to *b* size. For key sizes less than *n* bits are not recommended because it could weaken the cryptographic properties of the HMAC algorithm. Using a key *K* with a size greater than n bits does not add any cryptographic value to the algorithm. In cases in which the key's randomness characteristics must be increased, using a key larger than *n* is considered to be an advantage. (Keys longer than *b* bits are first hashed using *h*).

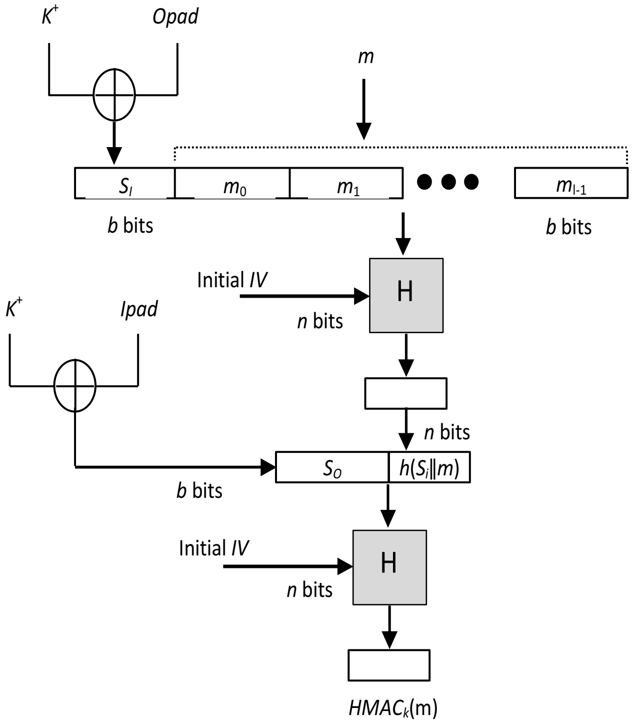


Figure 2. HMAC algorithm calculation

Furthermore, HMAC keys have two additional properties. First, the key must be random and calculated using an effective random value generation tool. Second, keys must be changed periodically, or a one-time key method must be used. There is no established guideline that specifies how many times a key can be repeated when guarding against attacks on HMAC and MAC algorithms. However, key changing is a cryptographic practice that is known to increase the strength of any cryptographic tool; such a practice will limit the exposure of keys [2].

HMAC functions can be compromised by one of the following schemes [9]. First, an attacker can compute hash results using a hash compression function, by using an exhaustive search attack (brute force attack) on the secret key in 2^n trials; second, the attacker can use a birthday attack to find two distinct messages m_1 and m_2 to calculate hash results $h(m_1)$ and $h(m_2)$, where $h(m_1) = h(m_2)$. A birthday attack requires $2^{n/2}$ trials to identify collisions for the selected hash function. In our case, we will require 2^{128} trials to break SHA-256. Although 2^{128} is a large number to calculate in logical time, the growing power of computers increases the threat that offline attacks will be used to find collisions. However, this attack can be effective only by guessing *K*. This can be accomplished by intercepting hash results calculated by the HMAC function, and using the same secret *K* to perform cryptographic attacks on these hash results.

IV. DESCRIPTION OF IMPROVED D-HMAC ALGORITHM

Improved d-HMAC [1] operates in a manner similar to HMAC, but uses dynamic values for *ipad* and *opad* instead of fixed values. The calculations for *ipad* and *opad* depend on three parameters: the content of message *m*, the *S*-box table, and the receiver's public key e_R. The main purpose of using these parameters in this work is to calculate different *ipad* and *opad* values for distinct messages *m*, and to calculate different *ipad* and *opad* values for distinct receivers, which are denoted by *R*. Improved d-HMAC is expected to be more secure than HMAC, because it utilizes dynamic *ipad* and *opad* values [12].

In this paper, we developed an algorithm to create and calculate *ipad* and *opad* dynamically with effective cryptographic properties. Several tests have been conducted to prove that the d-HMAC algorithm is superior to the HMAC function.

We will discuss two HMAC cases and explain how our improvements will solve the weaknesses presented below. These weaknesses make HMAC vulnerable to threats such as birthday and exhaustive attacks [13].

Case 1: Data *m* is sent to many recipients using the same content and the same secret key *K*, which results in a similar $\mathrm{HMAC}_K(m)$ for all recipients. This enables an attacker to collect message digests for message *m*, to generate offline attacks against HMAC. This weakness can be settled by sending distinct message digests $\mathrm{HMAC}_K(m)$ to recipients for the same data and key. Improved d-HMAC uses the public key e_R of the receivers, where every receiver has its own unique public key. This improvement will prevent an attacker from initiating an offline attack.

Case 2: Different messages are sent to the same recipient using the same secret key *K*. In this case, an attacker will collect messages *m* sent to the same recipient, knowing that they were calculated by HMAC with the same *K*, and subsequently attempt an attack. Obtaining some of the messages calculated by the same *K* will not break HMAC; however, it is a step toward increasing the attacker's ability to launch an attack. In d-HMAC, this weakness was solved by calculating the dynamic values of *ipad* and *opad* according to the message *m*; thus, the sender can safely send any number of messages using the same key *K*.

Algorithm 1 was developed to apply the ideas mentioned above, by calculating *ipad* and *opad* dynamically. It is depicted in Fig. 3. Message *m* and Public key e_R and *S*-Box (table T) in Table 1 are the input values for which the dynamic values of *ipad* and *opad* will be calculated.

Algorithm 1 ***(dynamic calculation for* ipad, opad *values)***

```
Input: message $m$, public key $e_R$, T.
Method:
   1. $w = h(m)$
   2. $ip = w \oplus e_R$
   3. $A=""$, $B=""$, $g=0$.
   4. for $i=1$ to $n/4$ do
        begin
            $g=g+1$
            $x=""$; $y=""$;
          for $j=0$ to 3 do
            begin
                $x = x \parallel ip[(g*i)+j]$
            end;
            $y = \overline{x}$
            $s = T[Decimal(x)]$;
            $z = Binary(s)$, where length of $z$ equals 4 bits
            $A = A \parallel z$
            $v = T[Decimal(y)]$;
            $w = Binary(v)$, where length of $w$ equals 4 bits
            $B = B \parallel w$
      end,
   5. $ipad' = A$, $opad' = B$.
Output: $ipad'$, $opad'$.
```

TABLE I. S-BOX (T)

3	2	1	0	
12	9	3	5	**0**
9	3	10	9	**1**
6	5	12	6	**2**
10	12	6	3	**3**

(a)

Value	
5	**0**
3	**1**
9	**2**
12	**3**
9	**4**
10	**5**
3	**6**
9	**7**
6	**8**
12	**9**
5	**10**
6	**11**
3	**12**
6	**13**
12	**14**
10	**15**

(b)

Note 1. To maintain the balancing property, the 4-bit values used in the *S*-box of in Table 1 all have two "0" bits and two "1" bits.

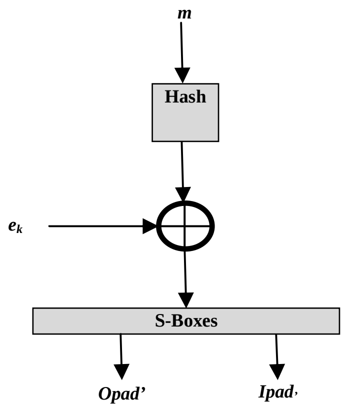


Figure 3. Dynamic *ipad'* and *opad'* generation

Algorithm 1 represents the methodology for calculating the dynamic values of *ipad'* and *opad'*. Such a calculation depends on three parameters in d-HMAC: the message *m*, the *S*-Box table, and the public key with length n from receiver e_R. Improved d-HMAC can use any type of hash function with different message digest lengths. Further, it can use any *S*-Box table with sufficient nonlinear characteristics. In this research, we focus on using the SHA-256 hash function with a 256-bit message digest, and on using table 1.

The following steps explain how *ipad'* and *opad'* are calculated in algorithm 1:

1. Calculate message digest $h(m)$ by using the SHA-256 hash function on *m*.
2. XOR message digest of message $h(m)$ with e_R.
3. Use *S*-Box to calculate *ipad'* and *opad'*.

The *S*-Box [16] was created to maintain balancing properties and nonlinearity. The *S*-Box's input and output both contain four bits. To explain *S*-Box calculations, we will review the following example, in which *S*-Box input = b_1, b_2, b_3, and b_4:

1. Concatenate $x = (b_1||b_2||b_3|| b_4)$.
2. Calculate *y*, where y is the complement of *x*.
3. Transform the 4-bit *x* into a decimal value from 0 to 15.
4. Select *s* and *v* from T, where $s = T[x]$ and $v = T[y]$.
5. Transform *s* to *z* and *v* to *w*, where *z* and *w* both contain 4 bits.
6. Concatenate *z* to *A* and concatenate *w* to *B*.
7. Repeat steps 1 through 6 to reach *n* bits.
8. Save the final value of *A* to *ipad* and *B* to *opad*.

In algorithm 1, we developed a calculation to preserve the maximum Hamming distance of *ipad* and *opad*.

The Hamming distance of two Boolean functions *f* and *g*, indicated by $d(f, g)$, can be calculated as follows:

$$d(f,g) = \sum_{x\in\{0,1\}^n} f(x) \oplus g(x).$$

Where $f, g\colon \sum n \rightarrow \sum$ are Boolean functions.

The Hamming distance for the two bit strings *ipad* and *opad* is calculated as following:

$$d(ipad',opad') = \sum_{x\in\{0,1\}^n} ipad' \oplus opad'$$

***S*-box**

S-Box values, which are used to calculate *ipad'* and *opad'*, are introduced in Table 1. The two-dimensional table T (a) consists of four columns and two rows. Each integer in the T table can be transformed into 4-bit binary values. We have selected values 3, 5, 6, 9, 10, and 12, which can be presented in binary as [0011], [0101], [0110], [1001], [1010], and [1100], respectively.

We considered two main cryptographic criteria in the construction of the *S*-Box:

- **Balancing property**: The values above fulfill the balancing cryptographic criteria [18,19]. The balancing property is an important cryptographic criterion, in which the number of ones and zeros in a string *S* with *n* bits will both equal *n*/2; in other words, the string contains an equal number of ones and zeros.
- **Hamming distance**: The distribution of values in table 1 was generated to maximize the Hamming distance of *ipad'* and *opad'*.

The d-HMAC function calculation mentioned in Fig. 4 is as follows:

$$d-HMAC_K(m) = h\big(K^+ \oplus Opad', h\big(K^+ \oplus Ipad', m\big)\big) \qquad (2)$$

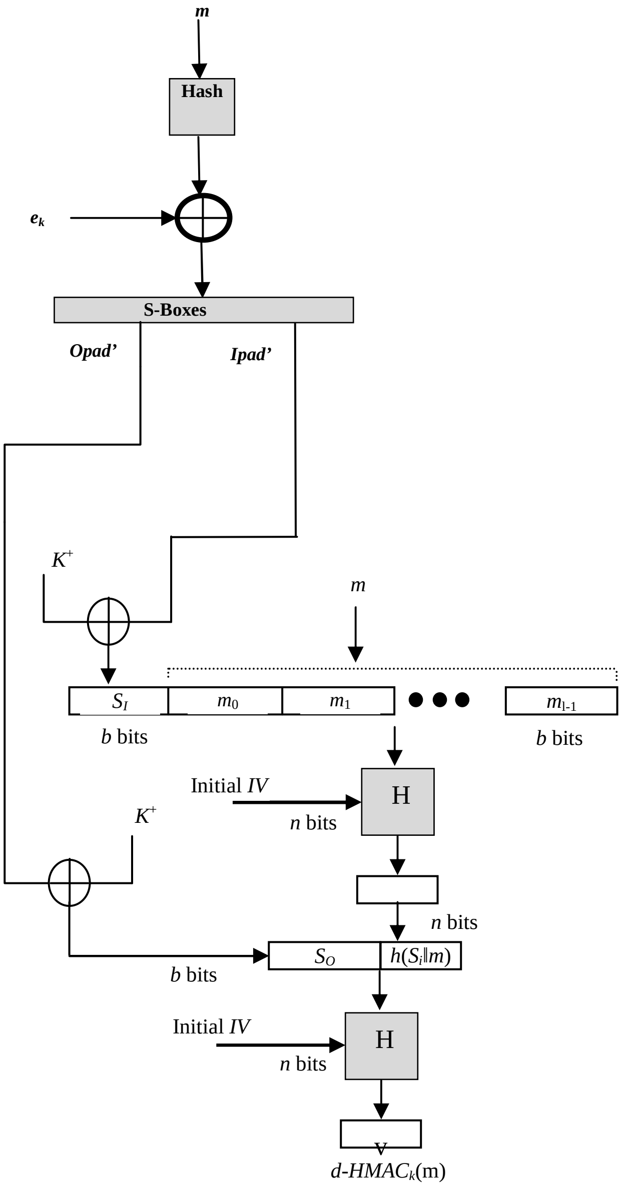


Figure 4. d-HMAC algorithm calculation

V. TEST RESULTS OF IMPROVED D-HMAC AGAINST HMAC

Improved d-HMAC is based on HMAC and it works in a similar manner, as mentioned in Section 3. These improvements are intended to enhance the security properties of HMAC, by increasing its resistance to different types of attacks. One of the advantages of improved d-HMAC is that it does not require any intermediate values to be stored, because these values are dynamically calculated for different recipients and different messages. Therefore, there is no need to implement a technique to protect these values from unauthorized disclosure.

Furthermore, improved d-HMAC does not require key frequent changes, because a random initial value is generated dynamically according to the recipient and the content of the message. The improved d-HMAC tests conducted in this study revealed that the modifications performed on HMAC to create d-HMAC had no negative impact on HMAC's main cryptographic characteristics; in some cases, it improved them.

Three types of tests were conducted. First, we compared the speed of improved d-HMAC against HMAC; in this test, both d-HMAC and HMAC used the SHA-256 algorithm. Second, we compared the algorithms' performance in terms of the avalanche effect. Third, we compared the balancing property [19]. All results are presented in tables and charts. The algorithms in the tests were implemented using C# 2010, and executed on an Intel Core i7 2.10 GHz processor running Windows 7 in 64-bit mode.

The Hamming distance method is used to calculate the avalanche effect and balancing properties by using XOR operations performed on the output bit string. The Hamming distance d for messages m_1 and m_2, in cases in which they are the same size n, can be calculated straightforwardly by counting ones in the set bits of $m_1 \oplus m_2$ and $0 < d < n$. The Hamming distance for m_1 and m_2 equals zero in cases in which $m_1 = m_2$, because there is no difference between m_1 and m_2 at the bit level. The Hamming distance for m_1 and m_2, where m_1 is a complement for m_2, equals n.

A. Speed test

As expected, d-HMAC required approximately twice the processing time as HMAC, because of the dynamic calculations for *ipad* and *opad* using the original message *m* (Roberts, 2008). These tests were conducted using files ranging in size from 1 MB to 100 MB; on our system, file sizes smaller than 1 MB generated processing times of approximately 62 ms and less for HMAC and 107 ms for d-HMAC, including hard drive access time (Table 2).

TABLE II. SPEED TEST FOR D-HMAC AND HMAC

File size (MB)	HMAC (s)	d-HMAC (s)
1	0.062	0.107
5	0.189	0.443
20	0.901	1.886
50	1.844	3.634
100	3.210	6.169

B. Avalanche effect test

The Avalanche effect is one of the primary design objectives for hash functions and any cryptographic tool; if an input message or secret key is changed slightly, approximately half of the output bits will be changed. Approximately 10,000 random sample inputs were tested for HMAC and d-HMAC; each avalanche effect test result listed in Table 3 represents a change of one input bit. In addition, 10,000 random key samples were tested for HMAC and d-HMAC; in these tests, one bit in the key was changed each time. The results are presented in Table 4.

TABLE III. AVALANCHE EFFECT RESULTS ACCORDING TO MESSAGE BIT CHANGING

XOR results	HMAC (bits)	d-HMAC (bits)
MIN	97	97
MAX	157	159
Average	128.003	127.956

TABLE IV. AVALANCHE EFFECT RESULTS ACCORDING TO KEY BIT CHANGING

XOR results	HMAC (bits)	d-HMAC (bits)
MIN	103	105
MAX	152	150
Average	127.775	128.295

C. Balance of ones and zeros test

Balancing is one of the primary design objectives for hash functions and cryptographic tools, in which the output bits contain approximately equal numbers of 1’s and 0’s. Ten thousand random input samples were tested for HMAC and d-HMAC; in these tests, one bit in the input string was changed each time (Table 5). Moreover, 10,000 random key samples were tested for HMAC and d-HMAC; in these tests, one bit in the key was changed each time. The test’s results do not indicate significant differences between d-HMAC and HMAC; see Table 6.

TABLE V. BALANCE RESULTS ACCORDING TO MESSAGE BIT CHANGING

XOR results	HMAC (bits)	d-HMAC (bits)
MIN	98	98
MAX	162	157
Average	127.962	128.062

TABLE VI. BALANCE RESULTS ACCORDING TO KEY BIT CHANGING

XOR results	HMAC (bits)	d-HMAC (bits)
MIN	106	106
MAX	149	149
Average	127.829	127.899

From the d-HMAC versus HMAC test results, we can conclude that the modifications required to create the improved d-HMAC preserved useful cryptographic properties; in some tests involving balancing and the avalanche effect, improved d-HMAC appears to exhibit better cryptographic characteristics than HMAC. The only difference is speed, where the improved d-HMAC requires approximately twice as much time as HMAC to perform calculations, which can even be a plus in password storage technology.

VI. D-HMAC RESISTANCE AGAINST CRYPTOGRAPHIC ATTACKS

One of the known attacks that can be used against HMAC is to collect many hash results that generate the same secret key *K*; this allows an attacker to initiate an offline attack by estimating the secret key. In contrast, it is infeasible to collect distinct messages calculated by the improved d-HMAC algorithm using the same secret key, because *ipad’* and *opad’* are always dynamically generated. As a result, d-HMAC is more resistant against cryptographic attacks than HMAC. Moreover, the HMAC function must securely store intermediate values [2], while improved d-HMAC does not require any cryptographic policy or tool to store any intermediate values ($K^+ \oplus$ *ipad’*) and ($K^+ \oplus$ *opad’*), as these values are dynamically calculated.

Programmers using HMAC for authentication must establish a secure method of saving intermediate values, which can be a serious vulnerability. In improved d-HMAC, there is no requirement to store these intermediate values, because they are dynamically calculated.

Finally, improved d-HMAC preserves useful cryptographic characteristics. The tests results show that improved d-HMAC did not compromise the avalanche effect or balancing properties.

VII. CONCLUSIONS

Our tests proved that using dynamically calculated *ipad* and *opad* values increases d-HMAC’s resistance to attacks [13], and does not negatively affect the avalanche effect or balancing. The dynamic *ipad* and *opad* values used in d-HMAC enable a sender to send as many messages as required to any number of recipients using the same key, and eliminate the possibility of an attacker collecting the message digests *ipad* and *opad* will be different, depending on the recipient. Furthermore, the sender can also send distinct messages to the same recipient using the same key, because *ipad* and *opad* values are changed with each calculation for these different messages. We also improved the *S*-box tables used in *ipad* and *opad* calculations by improving the cryptographic characteristics. Additionally, the tests showed that d-HMAC and HMAC have similar cryptographic characteristics.

We improved the d-HMAC function to be more resistant to brute-force attacks than HMAC. Additionally, the improved d-HMAC function uses *ipad* and *opad* values that are calculated dynamically, depending on several input parameters mentioned in this paper. This strategy can generate robust random strings that can be used in hash functions as initial values. As a result, SHA-256 and other hash functions having message digest sizes larger than 255 bits are more collision resistant to known attacks.

The developments and changes we implemented did not compromise cryptographic criteria such as the avalanche effect and balancing properties; this was proven by our tests described in Section 5.

In future work, we will expand the S-box table and improve its cryptographic characteristics. Furthermore, we will make the S-box dynamic, to more effectively control the algorithm’s calculation speed, which can be helpful to use d-HMAC in as a password storage function.Authors and Affiliations

REFERENCES

[1] Najjar, M. and Najjar, F., “d-HMAC Dynamic HMAC Function. Dependability of Computer Systems”, DepCos-RELCOMEX '06. International Conference on, 119-126, DOI: 10.1109/DEPCOS-RELCOMEX, 2006.

[2] Krawczyk, H., Bellare, M., Canetti, R., “HMAC: Keyed-Hashing for Message Authentication”, RFC 2104, 1997.

[3] Najjar, M., “Petra-r Cryptographic Hash Functions”, International Conference On Security and Management SAM ’03, Las Vegas, USA, Part I, 2003, 253-259.

[4] National Institute of Standards and Technology FIPS PUB 198, “The Keyed-Hash Message Authentication Code (HMAC)”, Federal Information Processing Standards Publication, 2002.

[5] Hansen, T., “US Secure Hash Algorithms (SHA and HMAC-SHA)”, RFC 4634, 2006.

[6] Tuner, S., Chen, L., “Updated Security Considerations for the MD5 Message-Digest and the HMAC-MD5 Algorithms”, RFC 6151, 2011.

[7] Wang, X., Yu, H., Wang, W., Zhang, H., and Zhan, T., “Cryptanalysis of HMAC/NMAC-MD5 and MD5-MAC”. LNCS 5479. Advances in Cryptology - EUROCRYPT2009, DOI: 10.1007/978-3-642-01001-9_7, 2009.

[8] Stokłosa J., “Bezpieczeństwo danych w systemach informatycznych (Data Security in Information Systems)”. Wydawnictwo Naukowe PWN, 2001.

[9] Stallings, W., “Cryptology and Network Security”, Prentice Hall, New Jersey, ISBN 0-13141098-9, 2011, pp: 362-394.

[10] Preneel B., “Cryptographic primitives for information authentication – State of the art. Preneel B., Rijmen V. (eds.)”, State of the Art in Applied Cryptography. LNCS 1528, Springer, Berlin, 1998, 49–104.

[11] Menezes A.J., van Oorschot P. C., Vanstone S.A., “Handbook of Applied Cryptography”. CRC Press, Boca Raton, FL, 1997.

[12] Speirs, W. R. (II.), “Dynamic Cryptographic Hash Functions”, UMI, 2007.

[13] Krawczyk, H., Bellare, M., Canetti, R., “Pseudorandom Functions Revisited: The Cascade Construction and its Concrete Security”, Proceedings of the 37th Symposium on Foundations of Computer Science, IEEE, 1996, 514–523, DOI: 10.1109/SFCS.1996.548510.

[14] Tsudik, G., “Message authentication with one-way hash functions, Proceedings”, INFOCOM’92, 1992, 22: 29–38, DOI: 10.1145/141809.141812.

[15] Schaad, J., Housley, R., “Wrapping a Hashed Message Authentication Code (HMAC) key with a Triple-Data Encryption Standard (DES) Key or an Advanced Encryption Standard (AES) Key”, RFC 3537, 2003.

[16] Hussain, I., Shah, T., Gondal, M. A., Khan, M., and Khan, W. A., “Construction of New S-box using a Linear Fractional Transformation”, World Applied Sciences Journal, 2011, 14 (12): 1779-1785, ISSN 1818-4952.

[17] ANSI x9.9, (revised 1986), American National Standard for Financial Institution Message Authentication (Wholesale) American Bankers Association, 1981.

[18] Olejar, D., Stanek, M., “On Cryptographic Properties of Random Boolean Functions”, Journal of Universal Computer Science, Springer, 1998, 4: 705–717.

[19] Lloyd, S., “Counting Binary functions with certain cryptographic properties”, Journal of Cryptology, 1992, 107–131, DOI: 10.1007/BF00193564.

[20] Roberts, D.,

http://www.powerbasic.com/support/pbforums/showthread.php?t=36862&highlight=d-hmac.

AUTHORS PROFILE

Mohannad Najjar received the B.S. and M.S. degrees in Computer Engineering from Poznan University of Technology, POLAND in 1998. Received Phd. In Telecommunication Engineering – Cryptography in 2002. During 2002-2009, he taught in Applied Science University- Amman, Jordan. Now he is the general manager on Vtech ltd. Company.

Hadoop Architecture and its Functionality

Dr. B V Ramana Murthy,
Department of CSE
Jyothishmathi College of Engg
and Technology, Shamirpet, India

Mr. V Padmakar
Department of CSE,
Guru Nanak Institutions
Technical Campus, Hyderabad

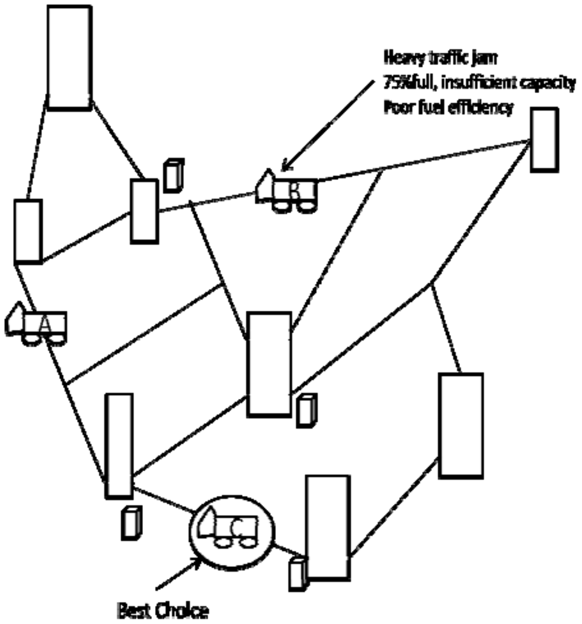
Mr. M Abhishek Reddy
Department of CSE,
Jyothishmathi College of Engg
Technology, Shamirpet, India.

Abstract:
Hadoop is nothing but a "framework of tools" and it is a java based programming framework (In simple terms it is not software). The main target of hadoop is to process the large data sets into smaller distributed computing. It is part of the Apache project sponsored by the Apache Software Foundation. As we observe in database management system, all the data are stored in organized form by following the rules like normalization , generalizations etc., and hadoop do not bother about the DBMS features as it stores large amount of data in servers. We are studying about Hadoop architecture and how big data is stored in servers by using this tools and the functionalities of Map Reduce and HDFS (Hadoop File System).

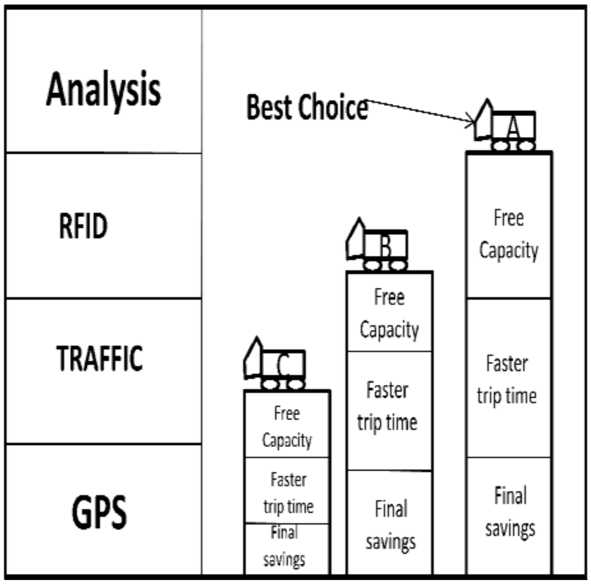
Keywords: Big Data, HDFS, Map Reduce Task Tracker, Job Tracker, Data Node, and Name Node.

Introduction-

Big Data Introduction: We probably heard about Big Data[1], but we may wondering what is it and why we should care. Ok, for starters big data is like data is getting bigger for a while now. From dawn of a time to less than a decade ago mankind generated about 5 Exabyte's of data. In 2012 global data brought to 2.7 Zetta byte of data which is 500 times more data than all data ever generated in 2003. And it has grown 3 times bigger in 2015. Some of the reasons that data is getting bigger[2] is that continuously being generated more sources and more devices. Much back data is like videos, photos, comments and social media comments on web sites is unstructured. That means data is stored in structures pre defines tables, instead it's often made up of volumes of text dates numbers in fact they are typically free from by nature. Certain data sources are arriving so fast not even a time to store the data before applying analytics to it. That is why traditional data management and analytics tools unable to store, process and analyze big data. So we could just ignore big data after all it's worth the efforts? Turns out, it is. A recent study concluded only 10-15% organization would take full advantage of big data. In order to generate that level of insight and competitive advantage from big data innovative new approach and technologies are required because big data we looking at is like a mountain.



Imagine a logistic company mining data on truck pickup and delivering schedule on real time traffic patterns. The data they are using combines real time GPS speed from trucks. Public traffic pattern data or if I take cargos from data. Imagine they get a call from a new pickup, which truck should they send? The closest one right. So what if the route to the closets truck has heavy traffic jam? What if the cargo loaded on that truck doesn’t allow space for new data? May be the route for that truck involve a series of great changes. In that case closest truck is not the best choice. They might b more costly less efficient or unable to service the customer needs. But the only way to arrival of optimal decision is to analyze multiple big data sources in real time.



Big Data = Big Impact

INTRODUCTION

As we see in our daily routine entire world has become an E-world (electronic world in common terms). So we can say increase in E-World in directly proportional to increase in data, so we required large no of servers to save the data.

To overcome this problem hadoop came into existence. Hadoop is so profound and powerful java tool which process large data into small data computations. Hadoop was created by Doug Cutting and Mike Cafarella in 2005 and Doug Cutting, who was working at Yahoo at the time named it after his son's toy elephant. And later they donated hadoop to apache so now we can say that hadoop is directed under the control of apache.

Hadoop architecture is playing a very important role in breaking of large data in to small data sets. In this paper we will know about architecture how the data[7] will get spitted and get computed and all its functionality.

Here comes a question in mind, how does Facebook, Google, Online marketing (retails), and all does store large amount of data?

The reason behind is all these frameworks uses hadoop system. The main reason hadoop came into existence of 3 factors.

They are velocity volume and verity. Velocity, large amount of data is coming with very high speed. Volume, large amount of data increasing day by day with huge volume. Verity, Data which are lots of verity. Ex: Audio, Video and etc.

Big data is creating large and growing files which are measured in terabytes (10^12) and petabytes (10^15) and the data is unstructured, we do not need relational models. This huge data is coming from tons of sources like users, applications like Facebook, yahoo, twitter etc, system, sensors and on and on.

The main problem hadoop is fixing is that in traditional hard disk transfer rate of data will be approx 60-100 MB/s and in hadoop there will be around 250-500 MB/s.

A. Reasons for hadoop evolution

Traditional Approach: when an enterprise will have a powerful computer it will process with very high speed it performance will be high and we can say computer is scalable. But there will be a certain point even a powerful computer cannot process Big Data. Now we can say computer is not scalable. This was one of the main reason hadoop came into existence.

Hadoop Approach: The main target of hadoop is to break Big Data[4] into smaller pieces and store into Commodity Hardware (Numerous Low Cost Computers known as Commodity Hardware). We do not require any powerful computers. At the same time all the computations are done on distributed system as well. All these computations are done at the same time and results send back to the application[6].

B. Hadoop simple Architecture

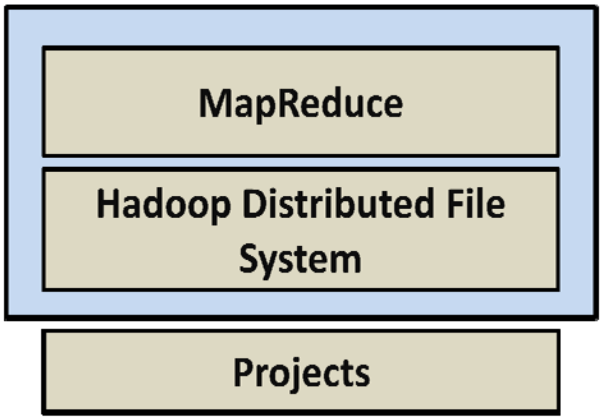


Figure 1: Hadoop Architecture

Hadoop Architecture consisting of three simple things i.e. MapReduce, HDFS, Projects. Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. Hadoop Distributed File System (HDFS) is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. Finally the projects, as we said hadoop is framework of tools so all the tools come under this project. Some examples for projects are Hive, HBase, Mahout, Pig, Oozie, Flume, Sqoop etc.

Hadoop consisting of two nodes:-

1. Slave node
2. Master node.

1. Slave Node: Slave node are having two major components

- Task Tracker.
- Data Node.

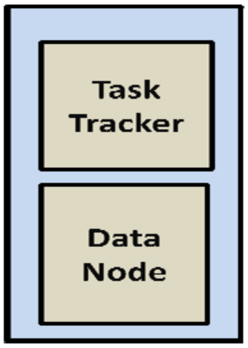


Figure 2: Slave node

1.1 Task Tracker: The job of task tracker is to processes the piece of task that has been given to this particular node.

1.2 Data Node: The job of data node is to manage piece of data that has been given to this particular node.

There can be n number of slave nodes. Here data is clustered in to these numerous slaves.

2. Master Node: The reason this is said to be a master node is that, master node having two another major components along with task tracker and data node.
1.1 Job Tracker.
1.2 Name Node.

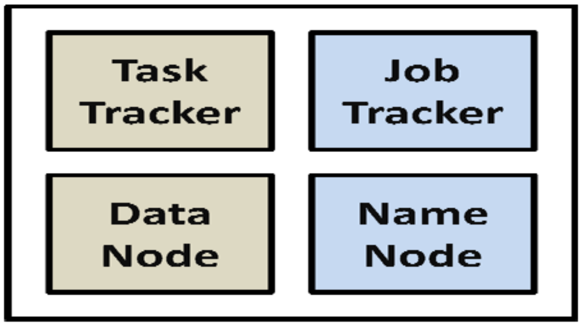


Figure 3: Master

1.1 Job Tracker: The role of job tracker component is to break higher task into smaller pieces and it will send each small computation to task trackers including its own. And after completing it will send back its results to the job tracker and it will combine the results and it will send back to application.

1.2 Name Node: It is responsible of keep an INDEX of which data is resigning on which data node.

Interaction between Master node and Slave node

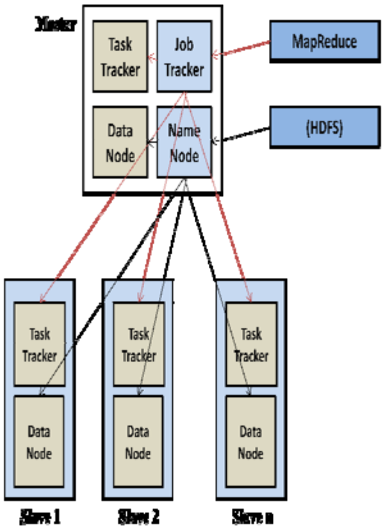


Figure 4: Interaction diagram

Job Tracker and Name Node functionality and interaction between them is observed in figure

MapReduce: Task Tracker and Job Tracker are the part of high level i.e. map reduce. So they all fall under the umbrella of map reduce.

File System: Data Node and Name Node are the part of high level i.e. map reduce. So they all fall under the umbrella of file system called HDFS.

Batch Processing:
One of the attribute of hadoop is that is a "Batch Processing" set of tools. So application would assign or provide a task for hadoop to in form of a QUEUE.

Once the task is completed it will inform application and results will be given back to application.

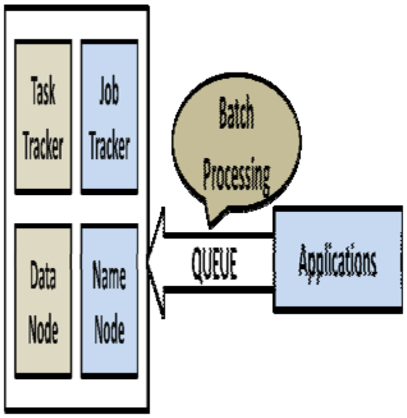


Figure 5: Batch Processing

Direction flow of data:

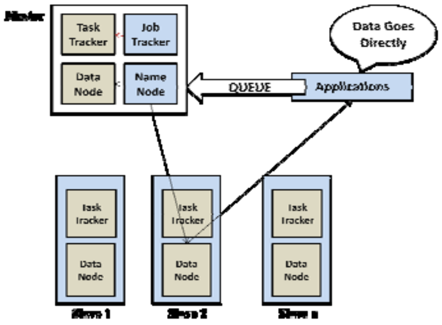


Figure 4: Data Flow Diagram

Here data flow directly when application comes in contact with master node and check the index on name node about the required information on which data node the data is residing. After the required information is gathered it directly goes to the application i.e. application doesn't wait for the name node to give back result. This is one of the important features is that; time optimizing in getting back result.

Fault Tolerance for Data and Master Backup:

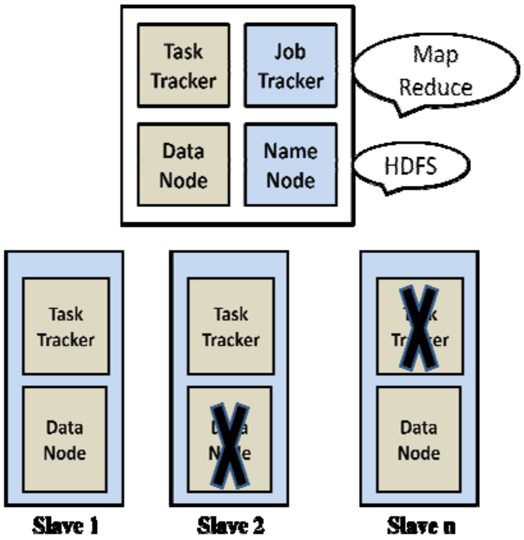


Figure 5: Fault Tolerance and Master Backup

One of the basic and important thing that hadoop keeps in mind is Fault Tolerance. If any of the Data Node gets failed, system doesn't go in to stop state by default Hadoop maintain 3 copies of each file and these copies are scattered along different computers. If any one of the task trackers gets failed to do its task, job tracker will detect the failure and it assigns the same task to other task tracker. When Master node gets failed then the tables that are maintained by name node which contain tables are backed up and copied over different computers. The enterprise version of hadoop also keeps two masters. One the main master and other the backup master.

Advantages of Hadoop:

One of the main advantages of hadoop to the programmers is

- Programmer need not worry about where the file is located name node will take care of it.
- Programmer need not worry about how to manage files; hadoop will take care of it.
- Programmer need not worry about how to break computations into pieces hadoop will take care of it.

- Programmer need not worry about writing the scalable programmers.

Consistency: - Component failures during execution of a job will not affect the outcome of the job.

Scalability: - Hadoop is highly scalable. As the no of slave nodes increases scalability also increases. Scalability of hadoop is linear, as we required processing speed to be increased then increase the no of computers.

Usage Areas:

There are tons of wide areas[3] where hadoop is used some of them are

- Social Media: Facebook, Twitter, yahoo, YouTube etc.
- Retail: e-Bay, Amazon etc.
- Searching Tools: Google.
- Companies: IBM etc.

And many more like American Airlines, The New York times and on and on. There are tons of users[5] who are using hadoop.

Conclusion

In this paper we have studied the entire architecture of hadoop and its functionality. It clearly explains that managing of big data in to clusters, how data is stored in numerous low cost computers (Commodity Hardware). Hadoop achieved Scalability and Consistency of data. As we seen in Database Management System we required organized data (following rows and columns) to store in server, we need follow normalizations techniques but where as in hadoop a programmer need not worry about relational data models.

Future Scope: - According to Yahoo point of view by the year 2015 50% of the enterprise[8] will processed by hadoop.

Biblography:-

[1] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.

[2] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov 18, 2011. http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[3] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Science 6 February 2009: 323 (5915), 721-723.

[4] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[5] Folowing the Breadcrumbs to Big Data Gold. Yuki Noguchi. National Public Radio, Nov. 29, 2011. http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-bigdata

[6] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data

[7] The Age of Big Data. Steve Lohr. New York Times, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[8] A Sustainable Future. Computing Community Consortium. Summer 2011.

[9] Windows Azure Platform: http://www.microsoft.com/windowsazure/

[10] Microsoft Service Bus and Access Control for Windows Azure platform AppFabric : http://www.microsoft.com/windowsazure/whitepapers/

[11] Windows Azure Tools – Constraints: http://msdn.microsoft.com/en-us/library/dd203058.aspx

[12] Microsoft Azure Comparison: http://cloudenterprise.info/2008/10/29/microsoft-azure-vs-amazon-google-and-vmware/

[13] Geneva Framework: http://download.microsoft.com/download/7/d/0/7d0b5166-6a8a-418a-addd-95ee9b046994/GenevaFrameworkWhitepaperForDevelopers.pdf

[14]SQL Azure: http://www.microsoft.com/windowsazure/sqlazure/

[15]WCF Data Services: http://msdn.microsoft.com/en-us/data/aa937697.aspx

[16]Windows Azure Platform AppFabric Services: http://msdn.microsoft.com/en-us/library/dd630576.aspx

Conclusion-

Windows Azure runs on machines in Microsoft data centers. Rather than providing software that Microsoft customers can install and run themselves on their own computers, Windows Azure is a service: Customers use it to run applications and store data on Internet-accessible machines owned by Microsoft. Those applications might provide services to businesses, to consumers, or both.

Dr. B. V.Ramana Murthy has done his PhD from Osmania University, presently he working as Professor in Computer Science and Engineering, has 18 years of experience in Teaching and R&D. His primary area of interest is Software Engineering & Web Engineering.

Mr. V Padmakar is pursuing PhD in CSE and has done his M Tech (CSE) from JNTUH, presently working as Professor in Computer Science and Engineering has 17 years of experience in Teaching and Industry. His primary area of interests is Software Engineering, Network Security and Data mining

Mr. Abhishek Reddy Mirupati is a Computer Science and Engineering student at Jyothishmathi College of Engineering and Technology pursuing his Bachelor of Technology in CSE. His primary area of interest is Object Oriented Programming and Data Base Management System.

Data mining methodologies to study student's academic performance using the C4.5 algorithm

Hythem Hashim[1], Ahmed A. Talab[2], Ali Satty[3], and Samani A. Talab[1]

[1] Faculty of Computer Science and Technology, Alneelain University, Khartoum, Sudan.
[2] White Nile College for Science and Technology, White Nile state ,Kosti.
[3] School of Statistics and Actuarial Sciences, Alneelain University, Khartoum, Sudan.

March 21, 2015

Abstract

The study placed a particular emphasis on the so called data mining algorithms, but focuses the bulk of attention on the C4.5 algorithm. Each educational institution, in general, aims to present a high quality of education. This depends upon predicting the unmotivated students before they entering in to final examination. Data mining techniques give many tasks that could be used to investigate the students performance. The main objective of this paper is to built a classification model that can be used to improve the students academic records in Faculty of Mathematical Science and Statistics. This model has been done using the C4.5 algorithm as it is a well-known, commonly used data mining technique. The importance of this study is that predicting student performance is useful in many different settings. Data from the previous students academic records in the faculty have been used to illustrate the considered algorithm in order to build our classification model.

Keywords: *Data mining*, *The C4.5 algorithm*, *Prediction*, *Classification algorithms*.

1 Introduction

The main objective of Faculty of Mathematical Science and Statistics in Alneelain University is to give quality education to its students and to improve the quality of managerial decisions. Thus, one recommendation is detect knowledge from educational records to study the main attributes that may affect the students performance in the considered faculty. This can be considered as an important and helpful aspect in choosing the right decisions to improve the quality of education. As well, it helps the academic planners in the faculty to enhance their decision making process with respect to the following aspects: (1) improve the student's performance; (2) improve teaching; (3) minimize the failure rate; and (4) other benefits. Data mining analysis is a good option to achieve the aforementioned objective as it gives many tasks that could be used to investigate the student performance.

When there is a need to data mining, the following question is forced upon researchers: What is data mining? In brief, the generic term *data mining* refers to extracting or ”mining” knowledge from large amounts of data. According to Suchita and Rajeswari (2013), it is a process of analyzing data from different perspectives and summarizing it into important information in order to identify hidden patterns from a large data set. The main functions of data mining are applying various techniques and algorithms in order to detect and extract patterns of a given stored data set (Jiawei et al., 2012). There are several studies in the literature which provide a comprehensive review of the data mining applications. For instances, see Florin, G. (2011) and Jiawei et al. (2012). According to Barros and Verdejo (2000) and Jiawei et al. (2012), data mining can be classified into various algorithms and techniques, such as classification, clustering, regression, association rules, etc, which are used for knowledge discovery from databases. These techniques should be fully understood and appropriately characterized in relation to educational data analysis and should be theoretically proved before they are used practically. A brief overview of some of these techniques is given in the next section. For more detailed discussion of data mining, see Michael and Gordon (2004) as well as Ian and Eibe (2005). Here, we have to make it clear that, this research restricts attention to only consider the classification task for assessing student's performance, specifically the C4.5 algorithm is the main focus of this paper.

In this paper, student's information, such as their degrees in the previous academic records (annually) are collected to predict the performance at the end of the last year based on various attributes. The study was done on the data set that has 124 graduate students. Further, we have identified the important and necessary attributes that impact the student's academic performance. An application study is carried out using the Weka software and real time data set available in the college premises. The paper aims to predict the student's performance in the faculty result based on the basis of his/her performance throughout the study period. The paper is organized as follows: In Section 2, a background for data mining is provided, followed by a particular focus on the decision trees modeling based on the C4.5 algorithm. Section 3 consists of our application schemes study, including a description of the data set used in the analysis. The results are then described and discussed in Section 4. The study concludes in Section 5, with a brief description of some concluding remarks.

2 Data mining algorithms

As stated in Ian and Eibe (2005), data mining algorithms have become a huge technology system after years of development. Generally, data mining has the following basic topics: **(1)** Classes: stored data are used to locate objects in predetermined groups; **(2)** Clusters: data items are grouped according to logical relationships or consumer preferences; **(3)** Associations analysis: data can be mined to identify associations; **(4)** Sequential patterns: data is mined to anticipate behavior patterns and trends; and **(5)** Prediction: data can be used to find out their trends and behavior and to predict their future behavior according to the historical data stored in data warehouse. As discussed earlier, the classification task is a focus of this article. Sun et al. (2008) stated that classification is a systematic technique based upon the input data to establish a classification model. Moreover, the classification examples consists of the following algorithms: decision tree, rule-based, Naiive bayes, etc. However, despite these number of classification methods, we focus on the C4.5 algorithm which is one of the decision trees classification algorithm as it has been the data mining approach of choice. The main aim of classification is to build a model in training data set in order to predict the class of future objects whose class label is not known. There are two broad topics in classification; these are: **(1)** Preparing the data for classification and prediction; and **(2)** Comparing classification and prediction methods (Ian and Eibe, 2005). Furthermore, classification employs a set of pre-classified examples to develop a model that can classify the population of records at large. In general, the data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In addition, the classification test data can be applied to estimate the accuracy of the classification rules (Florin, 2011).

2.1 Decision trees modeling - the C4.5 algorithm

Decision tree modeling is one of the classifying and predicting data mining techniques, belonging to inductive learning and supervised knowledge mining. It is a tree-diagram-based method, depending on two manners; the node on the top of its tree structure is a root node, and nodes in the bottom are leaf nodes. Target class attribute is given to each leaf node. From root node to every leaf node, there is a path made of multiple internal nodes with attributes. This path generates rule required for classifying unknown data. Moreover, most of decision tree algorithms contain two-stage task, i.e., tree building and tree pruning. In tree building stage, a decision tree algorithm can use its unique approach (function) to select the best attribute, so as to split training data set. The final situation of this stage will be that data contained in the split training subset belong to only one certain target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, some special data in training data set may lead to improper branch on decision tree structure, which is called over-fitting. Therefore, after building a decision tree, it has to be pruned to remove improper branches, so as to enhance decision tree model accuracy in predicting new data. Among developed decision tree algorithms, the commonly used ones include ID3, C4.5, CART and CHAID. The C4.5 algorithm is an extension of the ID3 (Iterative Dichotomiser 3, it is a simple decision tree learning algorithm developed by Quinlan (1986)) algorithm, it uses information theory and inductive learning method to construct decision tree. C4.5 improves ID3, which cannot process continuous numeric problem. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. Further details of these algorithms can be found in Kass, G. V. (1980), Ian and Eibe (2005) and Sun et al. (2008). Decision trees based on the C4.5 algorithm is a commonly used classification techniques which extract relevant relationship in the data. The C4.5 algorithm is a program that creates a decision tree based on a set of labeled input data. Further, the decision trees modeling created by this algorithm can be used for classification, and for this reason, the C4.5 algorithm is often defined as a statistical classifier. The C4.5 algorithm makes decision trees using a set of training data, taking into account the concept of information entropy. The training data can be defined as a set $S = s_1, s_2, ...,$ of already classified samples. Thereafter, each sample $S_i = x_1, x_2, ...$ is a vector, where $x_1, x_2, ...$ denotes attributes of the sample. Then the training data is augmented with a vector $C = c_1, c_2, ...$ denotes the class to which each sample belongs.

3 Application study

3.1 Data description

In this paper, we consider student's data set that are pursuing Bachelor of Statistics, Actuarial Science degree from Faculty of Mathematical Sciences and Statistics in Alneelain University. The variables used for assessing the student's performance as well as for building a predicted models in the faculty were degree1, degree2, degree3, degree4 and degree5, corresponding to the students degrees in the period from 2008 to 2013. The number of graduates selected was 124. As discussed earlier, the study was focussed on the previous academic records of the students. The first fourth student degrees have been used in order to predict the student degrees with respect to the fifth degree. Of these 124 records in our data set, we set up that each record has five numerical attributes. These attributes provide the annual degrees in which values ranged from 1 to 124. The data has been preprocessing in three stages:

1. Convert the first fourth degrees into nominal data type according to the following syntax:

```
  if in [40..59]:Pass,
    if in [60..69]:Good,
      if in [70..79]:V.Good
otherwise:Excellent
```

2. Handling the missing attribute information using the imputation technique. A useful discussion in terms of the technical details of the imputation technique is given by Satty and Mwambi (2012)

3. Divide the class label into the three broad classes. This has been done using the following syntax:

```
if less than 59: class C (students need extreme improvement to their degrees)
  if in [60..79]: class B (students need a little bit improvement in their performance)
    otherwise: class A (this class includes those students who were doing well)
```

3.2 The WEKA software

According to Remco et al. (2012), WEKA is defined as an open source application that is freely available under the GNU general public license agreement. Firstly, this software has been originally written in C, the WEKA application thereafter has been completely re-written in Java, and is compatible with almost every computing platform. Generally speaking, it is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. It can be used to apply many different data mining tasks such as data preprocessing, classification, clustering, and so on. However, in this paper, we only placed a particular emphasis on considering the C4.5 algorithm as it is a commonly used classification algorithm. More details of WEKA, including its characteristic system, file format, system interface, the mining process can be found in Remco et al. (2012). It accepts the data in specific formats, such as ARFF (Attribute-Relation File Format (ARFF), CSV (Comma Separated Values) and C4.5.s format. These specific formats have been taken into account in dealing with this paper.

3.3 Fitting a C4.5 algorithm

The main objective of fitting this algorithm is to provide a model that accurately predicts the class of the unknown tuples or records. To do so, we used the following steps that represent basic principle of working for this classifier: **(1)** We provide the training set that consists of the training records along with their associated class label; **(2)** We build the classification model by applying the learning algorithm used in respective technique; and **(3)** The model built is applied on the test set that consists of the tuples that do not have the associated class label. This algorithm has been carried out using the CRISP process. CRISP refers to CRoss Industry Standard Process, which contains six stages. Figure 2 displays the link between them.

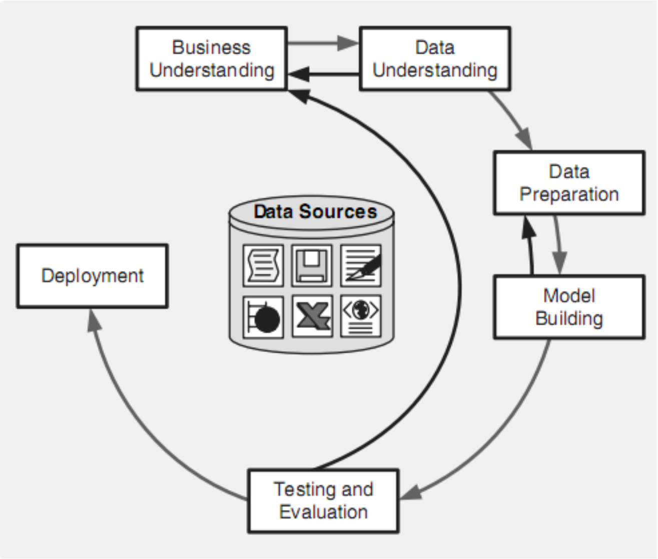


Figure 1: CRISP process

3.4 Measures for performance assessment

For a binary decision problem, a classifier labels examples as either positive or negative. The confusion matrix or can be used constructed to make the decision that can be made by classifier.

This matrix consists of four categories: True positives (TP) are examples correctly labeled as positives, ; false positives (FP) correspond to negative examples incorrectly labeled as positive; true negatives (TN) reger to negative correctly labeled as negative; and false negatives (FN) correspond to positive examples incorrectly labeled as negative. We further define TPR as the true positives rate, which is equivalent to Recall (would be briefly visited below). This matrix builds the so-called recall and precision measures. Recall can be computed as Recall = TP/(TP+FN). Precision measures that fraction of examples classified as positive that are truly positive. It can be calculated as Precision = TP/(TP+FP). In fact, there are several basic measures that can be used to assess the student's performance. Such measures are readily usable for the evaluation of any binary classifier. Consequently, to assess the performance of the data set mentioned above, we use these evaluation criterions depending on the next measures: **(1)** Accuracy: it is defined as the number of correct predictions divided by the total number of predictions; and **(2)** Error rate: It refers to the number of wrong predictions divided by the total number of predictions. Furthermore, the description of each measure here is shown below: The correctly classified instances show degree of test instances that were correctly classified (Accuracy). The incorrectly classified instances show age of test instances that were incorrectly classified (Error Rate); **(3)** The Kappa statistic: It was introduced by Cohen (1960). Bartko and Carpenter (1976) has sated that Kappa is defined as a normalized statistic measure of agreement that is computed by taking the agreement expected by chance away from the observed agreement between the classifier and actual truth and dividing by the maximum possible agreement. The possible value for Kappa lies in the range [-1, 1] although this statistics usually falls between 0 and 1. The value of 1 indicates perfect agreement, however, the value of 0 indicates that the agreement no better than expected by chance. So, when the value of k is greater than 0, it implies that the classifier is doing better compared to chance thus indicating perfect agreement at $K = 1$ else if the value of k is 0, then it denotes the chance agreement. A kappa with the negative rating indicates worse agreement than that expected by chance. Now, let P_a and P_e denote the percentage agreement and expected chance (hypothetical) agreement, respectively. Thus the Kappa statistic can be expressed as follows: $K = (P_a - P_e)/(1 - P_a)$. In our analysis, for computing k, we have the total instances = 124. **(4)** F-Measure combines recall and precision scores into a single measure of performance. It can be computed as F-Measure=2*(recall*precision)/(recall + precision). **(5)** ROC area (Receiver Operator Characteristic) is commonly used to provide findings for binary decision problems in data mining. Using it together with the recall and precision measures we can get a more informative picture of the C4.5 algorithm.

4 Results and discussion

From the C4.5 classification algorithm, the decision tree is constructed, depending on the most effective attribute(s) is/are given using the so called Entropy and the Gain information. Hence, to achieve this construction, we needed to calculate the entropy of each features of the training images by using C4.5 algorithm and measure the information gained for each features and take maximum of them to be the root (Andreas and Zantinge, 1996). $Entropy(S) = \sum_{i=1}^{n} -P_i log_2 P_i$, and the Information Gain is given by $Gain(S, A) = Entropy(S) - \sum \frac{|S_\nu|}{|S|} Entropy(S_\nu)$, where where P_i is the probability of a system being in cell i of its phase space, $\sum_{i=1}^{n} -P_i log_2 P_i$ gives the entropy of the set of probabilities $P_1, P_2, ..., P_n$, $\sum$ is over each value ν of all the possible values of the

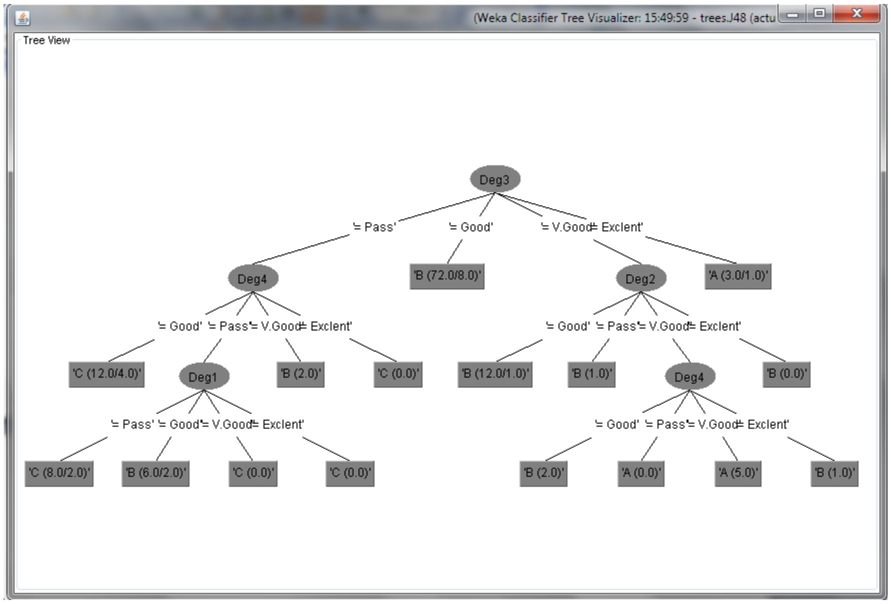


Figure 2: Decision tree for students degrees data set

attribute A, ν is the subset of S for which attribute A has value ν, $\mid S_\nu \mid$ is the number of element in S_ν, $\mid S \mid$ is the number of element in S. Anyhow this paper is interested in finding out the relationships between the considered degrees attributes. Therefore, decision tree is displayed in Figure 1. This figure shows that, depending upon the information gain, the attribute deg3 has the maximum gain and hence it comes at the top of the decision tree (decision tree algorithms use the gain value to start splitting the tree with attribute having high gain and so on). The number of leafs from the decision tree output was 16. The tree size that is obtained was 21, the time taken to build our model was 0.02 seconds. Moreover, The partial tree below is the result of applying the C4.5 classification algorithm model, in which the tree consists of 5 leaves marked with L_1, L_2, L_3, L_4 and L_5.

1. Deg3 = pass
2. | Deg4 = Good: C (12.0/4.0)
3. | Deg4 = pass
4. | | Deg1 = pass: C (8.0/2.0)
5. | | Deg1 = Good: B (6.0/2.0)
6. | Deg4 = V.Good: B (2.0)
7. Deg3 = Good: B (72.0/8.0)

The leaf L_1 contains instances (12, 4) in the row number 2, node Deg4. Therefore, in this leaf there were 16 records from the data set have classified in class C. The leaf L_2 contains instances (8, 2), which is to say that 10 records were classified in class C. L_3 contains instances (6, 2) in row number 5, Deg1 = Good, this implies that 8 records have been classified in class B. L_4 consists of instances (2) in row number 6, node Deg4 = V.Good, which means that this leaf has 2 records that have classified in class B. Finally, L_5 consists of instances (72, 8) in row number 7, node Deg3 = Good, which means that 80 records were classified in class B. The results further yielded the confusion matrix. From this matrix we extract the following findings: (1) 25 records were classified in class C, thus 20% belong to C, 14 of these records ere TP with rate of 56%; (2) 91 records have been categorized to be in class B. This refers to that 73% belong to B, and 77 records were TP with the rate of 85%; and (3) 8 records have been seen in class A, meaning 6% belong to this class. 5 records of them were TP under the rate of 63%.

The results for computing the accuracy and error rate measures are displayed in Table 1. By looking at this table, we find that (as the number of instances was equaled to 106) the accuracy $= (106/124) * 100 = 85.4839\%$, and the number of incorrectly classified equals 18, the error rate therefore $= (18/124) * 100 = 14.5161\%$. As wee see in Table 1, in order to fit the C4.5 algorithm, we provide the training set to build a predictive model. This training set consists of the predictor attributes as well as the prediction (class label) attribute. First, we use the training set in the preprocess panel, followed by the selection of the C4.5 algorithm. The 10 fold cross validation option is being selected. Second, we apply the same procedure on our testing set to check what it predicts on the unseen data. For that, we select "supplied test set" and choose the testing data set that we created. Finally, we run the C4.5 again and we notice the differences in accuracy. Note that the correctly/incorrectly classified instances define the case where the instances are used as test data. Depending on these findings, we see that 85.4839% can be considered as a good percentage to achieve the main goal of this paper. Turning to the error rates displayed in the table, we see that the error rates are the same for both training and supplied tests. This indicates that the considered algorithm was doing well for both tests. However, for cross validation folds as well as for percentage split 66%, the error rates were different, which is to say that C4.5 was effective with respect to cross validation fold as it has a lower error. This can be justified by the fact that an algorithm will be preferred when it has a lower error rate, namely it has more powerful classification capability and ability in terms of student's performance. On the other hand, the Kappa statistic that we obtained was 0.6327, which is to say that the used algorithm in our model is well doing as the Kappa statistics is greater than 0 (see, Cohen, 1960 in terms of interpreting a Kappa statistic). The results of recall, precision, F-measure and ROC area are displayed in Table 2. The findings yield that the recall and precision present estimates closer to each other. Note that in precision and recall measures, as the level of recall varies, the precision does not necessarily change linearly because of the fact that FP replaces FN in the denominator of the precision matric. As we know higher precision as well as F-measure are better. Thus, as given in Table 2, the the findings were high (above 70%) leading to that fact that the C4.5 algorithm is an effective and reliable technique to be recommended.

Table 1: Testing options

Training option	Correct classify instance %	Incorrect classify instance %
Training set	85.4839%	14.5161%
Supplied test	85.4839%	14.5161%
Cross validation folds = 10	77.4194%	22.5806%
Percentage split 66%	76.1905%	23.8095%
Kapa = 0.6327		

Table 2: Detailed accuracy for each class - classification using the C4.5 algorithm

	TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
	0.56	0.061	0.7	0.56	0.622	0.832	C
	0.934	0.333	0.885	0.934	0.909	0.844	B
	0.875	0.009	0.875	0.875	0.875	0.992	A
Weighted avg.	0.855	0.257	0.847	0.855	0.849	0.851	

5 Conclusion

In this study, we have placed a particular emphasis on the so called data mining algorithms, but focuses the bulk of attention on the C4.5 algorithm. Our goal was to build a predicted model that can be used to improve the student's academic performance. In order to achieve this goal, data from the previous students academic records in the faculty have been used to illustrate the considered algorithm in order to build our predicted model. Despite there are other classification algorithms, the C4.5 approach has been the principal data mining algorithm of choice for the primary analysis for dealing with students performance prediction because of its simplicity as well as the ease with which it can be implemented. Here we refer to statistical softwares such as, SPSS and SAS. Thus, the C4.5 approach might become attractive in specific circumstances. We here believe that the C4.5 algorithm can be recommended as a default tool for mining analysis. The findings in general revealed that it is possible to predict the probability of getting a degree within the estimated period according the degree of a graduate in the attributes performance. In conclusion we submit that the algorithm described here can be very helpful and efficient if there is an application study regarding the assessment of students performance, where both kind of knowledge is required (association among attributes and classification of objects).

References

[1] Andreas, P. and Zantinge, D. (1996). Data mining, Addison-Wesley, New York.

[2] Barros, B. and Verdejo, M. F. (2000). Analysing student interaction processes in order to improve collaboration: the degree approach. *International Journal of Artificial Intelligence in Education*, 11, 221-241.

[3] Bartko,J.J. and Carpenter, W.T., (1976). On the methods and theory of reliability. *J NervMent Dis*, 163, 307-317.

[4] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

[5] Florin, G. (2011). Data mining: concepts, models and techniques. Springer-Verlag. Berlin Heidelberg.

[6] Ian, H. W and Eibe, F. (2005). Data Mining: practical machine learning tools and techniques, Second Edition. Elsevier Inc. San Francisco: USA.

[7] Jiawei, H., Micheline, K. and Jian P. (2012). Data mining concepts and techniques, Third edition. Elsevier Inc: USA.

[8] Kalyani, G. and Jaya, A. LakshmiPerformance assessment of different classification techniques for intrusion detection. *Journal of Computer Engineering (IOSRJCE)*

[9] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

[10] Michael, J.A. B. and Gordon, S. L. (2004). Data mining techniques for marketing, sales, and customer relationship management, Second edition. Wiley Publishing, Inc. Indianapolis, Indiana: USA.

[11] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81-106.

[12] Remco, R., Eibe, F., Richard, K., Mark, H., Peter, R., Alex, S. and David, S. (2012). WAIKATO, Weka manual for version 3-6-8. Hamilton: New Zeland.

[13] Satty, A. and Mwambi, H. (2012). Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, 46, 327-356.

[14] Suchita, B. and Rajeswari, K. (2013). Predicting students academic performance using education data mining. *International Journal of Computer Science and Mobile Computing*, 2, 273-279.

[15] Sun, G., Liu, J. and Zhao, L. (2008). Clustering algorithm research. *Software Journal*, 19, 48-61.

Resource modeling for the development of a decision-making system - applied HCEFLCD of Morocco

K.OUBEDDA, M.KHALFAOUI, A.ETTAHIR

Systems Analysis Laboratory of Information Processing and Integrated Management (LASTIMI) School of Sale Technology, University Mohammed V Agdal

ABSTRACT. The aim of this work and to develop a decision support system for the operation of a model including the main stakeholders of the High Commissioner for Water, Forests and Desertification Control (maker / managers, administrative, customers). This system is based on the relationship between the actors and their activities and their needs vary by contribution in time. It aims to make available to managers a set of dashboards that can improve the quality of provided services. We begin by modeling the actors up and clean process for studying both their organizations and their activities and needs. The first applications of this work has focused on data for the Directorate of Planning, Information System and Cooperation, and the Directorate of Forest Estate, Legal Affairs and Litigation. The results are encouraging.

KEYWORDS

Information systems, decision support systems, dashboards, databases, modeling.

1- Introduction

Business intelligence is one of the areas of computer knows nowadays a copy boom. Indeed, business managers, faced with increasingly unstable environments, are expected to take the most effective decisions based on reliable data. The current problem is not to have a better decision tool, but to structure upstream data that feed not to be ineffective. Thus, the design of decision-making information systems tailored and scalable is a topical issue for all organizations around the world. Today HCWFDC assigned missions require not only to be responsive to the needs of its users / decision maker but also to anticipate these needs by acting as a "business". Also, BI is becoming a necessity for the control system of the Moroccan HCWFDC.

This work is in this context and focuses on the design and implementation of a decision support system dedicated to HCWFDC [7]. The goal is the modeling of the actors that reflects their business and their needs. In doing so, we aim to provide as complete and accurate as possible a description of all aspects of the behavior of the actors and the system able to provide dashboards to facilitate decision making. [8] The adopted model reflects a true picture of the system and uses UML current standard in the world of design information systems.

2- Hypothesis

We begin by modeling [4] actors up taking into account the requirements and expectations of each of them, namely for an administrative actor:
- Needs to have good training with skills facilitating their integration into working life.
- The activities and to facilitate and serve the work of the decision maker; and to disseminate and share information, meet the needs of external customers; ...

Given this situation, it is to correlate between the needs of decision makers [1], [2]; of the administrative. In fact, we are faced with a situation of looking for the satisfaction of the customer / user with a specificity High Commission of the notion "actor / user" Also as a business. Indeed, for the corporate governance approach is for profit, while for a public agency HCWFDC, it is more about positioning and visibility of the organization. The company seeks a performance positioning in its capital and HCWFDC aims to achieve a quality label and better manage natural resources. The company seeks customer satisfaction, HCWFDC seeks to satisfy its players. Customer satisfaction in business is formalized in terms of costs. Satisfaction actors HCWFDC is illustrated by satisfying their information needs at the right times.

3- Context

Today High Commission for Water, Forests and Combat Desertification Moroccan stirs in its information system a large volume of data and information. [2] Often, the very fact of this volume, it becomes difficult to make sense of these data and exit accurate and reliable indicators. To exploit these data, ensure and automate the management of criminal litigation, policy makers at the central level do not have homogenized data. Needs have deployed a final version of the application at national level (with the exception of the South HCWFDC). Thus, we group our data warehouse model into three levels:

Actor level: Decider two classes, Administrative (Head of Litigation CCDRF, DPEFLCD ....)
Repository -Database: includes all the data (Region Natural Resource available ... ..)

Administrative level: Database on the administrative status of all players.
Basic regulations: Regulations and Rights of each actor.

4. Modeling of players

In previous work [1] [2], we showed that applications for player level based on information which the data are collected in databases available to the central management and in the regions (dates, location offenders, nature ...). The design of decision support system [3] requires special design approach and complex modeling. [4] It has adopted a model to meet specific needs such as factor analysis [5] that the principle of facilitating the understanding and interpretation of a large set of multidimensional data. This analysis graphically highlight similarities between data and quantifies the degree of correlation between several factors.

We obtain the model that includes all the actors involved in the following university system:

$$\text{Acteur=} A \; ; \; \sum_{j=1}^{j=3} R_j \; ; \; \sum_{k=1}^{k=n-1} RC_k \; ; \qquad (1)$$

Along With:

A: Regions distinguishes three regions.

RC: Claims varies by contribution time differs: Module list pvs, pvs Details Module, Module monitoring judgments,

After the codification of Formula 1, we get:

The portfolio of the source (S) defined all the functional needs of the activities to be carried out over time by each region.

❖ Building Multidimensional conceptual model:
UML formalism is used to model the types of actors (decision maker, administrative). Indeed, UML has a graphical notation visual form based on charts that can facilitate decision making.

5- Visualization Data

To assist in the representation of our model is visualized as an example the data around a mass of complaints by contribution to natural resource available for each region. The figure below shows the actual activities of registered PVS claims that actor in relation to the information system during the year.

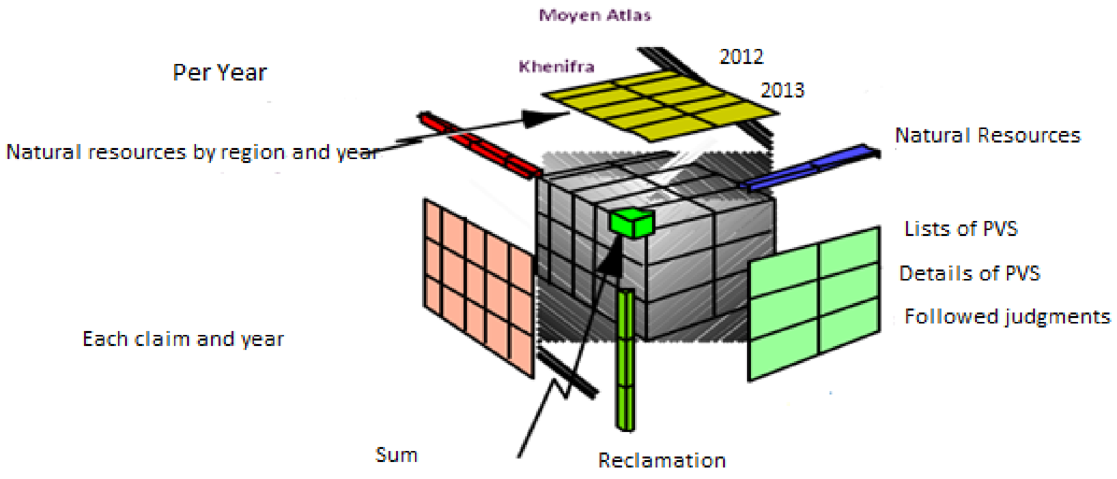
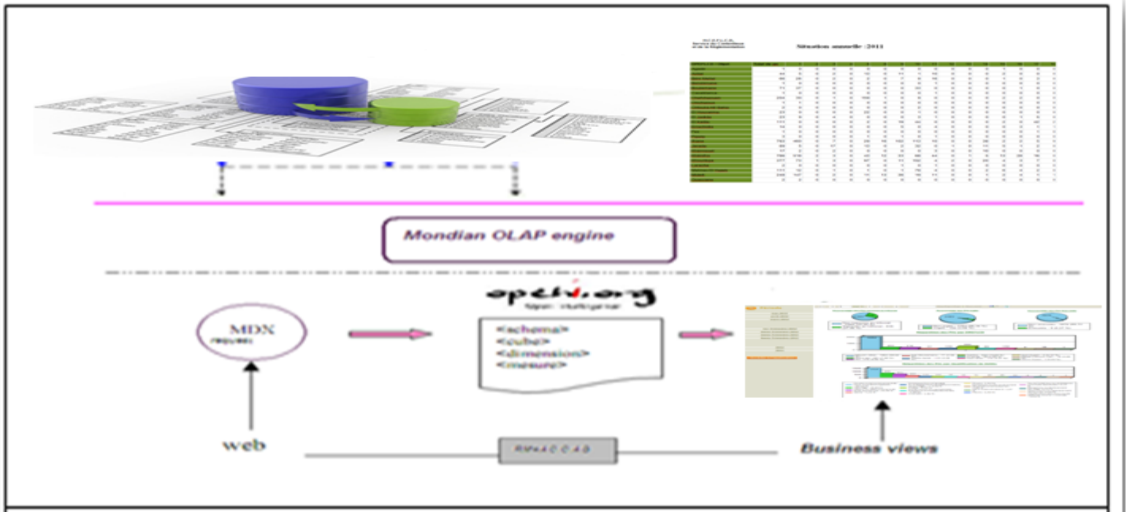


Figure 2: Vision of Claims and Natural Resource by region and for each year

6- Expected impacts on flying the HCWFDC

We present below the form of our model with all parameters. It is to inform decision makers in the HCEFLCD and taken as scope.



XMLA: ExtensibleMarkup Language Analysis

MDX: MultiDimensional eXpression.

Mondrian: Server OLAP written in Java.

Openi: GUI based on Mondrian.

Conclusion

To install this application, we went through three major phases. The first concerns the theoretical part that needs to have a model that can respond to the context of our environment known by its complexity (actors, data resources, non-homogeneous data, ...). This required a simple mathematical model defining the relationships between the actors, their activities and their needs. The second phase focused on the consolidation of data and designing a multidimensional database. The third phase was devoted to the application and to build a dashboard checking all the proposals made in the theoretical part.
The availability of real data of other players would be enough to have a global decision-making tool for HCWFDC.

Bibliography

1. Oubedda L, erraha. A, Khalfaoui. M, " Tools for decision support in planning academic needs of actors" ,IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012
2. Oubedda L, erraha. A, Khalfaoui. M, " Decision-making application for the Management of Human Resources: the automation of the recruitment to the breasts of Universities" ,IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012
3. Oubedda L, erraha. A, Khalfaoui. M «Conceptual modeling in the ontological basis of a Data Warehouse- Environment University" ,IJCSIS International Journal of Computer Science and Information Security, Vol. 10, No. 2, 2012

4 .Olivier Bistorin, Thiband Monteiro, Claude Pourcel, "Modélisation des processes d'un system de formation", Proceedings 1ère Conférence Internationale sur l'Ingénierie des System de Formation, Carthagéne des Andes, Colombie, octobre 2007.

5 Peguiron F,David A,Thiery O, "Intelligence économique dans un cadre universitaire intégrant la modélisation de l'utilisateur", IERA 2003,Nancy.

6. Bouaka N. et David A., «Modèle pour l'explicitation d'un problème décisionnel : un outil d'aide à la décision dans un contexte d'intelligence économique", IERA 2003, Nancy.

7. THIERY O., DUCREAU A., BOUAKA N., DAVID A., "Piloter une organisation : de l'information stratégique à la modélisation de l'utilisateur ; application au domaine de la GRH", GREFIGE, 2004.

8. Robin Jaulmes — Joelle Pineau —Doina Precup, "Apprentissage actif dans les processus décisionnels de Markov partiellement observables_L'algorithme MEDUSA",

IJCSIS REVIEWERS' LIST

Dr. Mahmoud M. A. Abd Ellatif, Mansoura University, Egypt

Prof. Kunwar S. Vaisla, BCT Kumaon Engineering College, India

Prof. Mahesh H. Panchal, Kalol Institute of Technology & Research Centre, India

Mr. Muhammad Asad, Technical University of Munich, Germany

Mr. AliReza Shams Shafigh, Azad Islamic university, Iran

Prof. S. V. Nagaraj, RMK Engineering College, India

Mr. Ashikali M Hasan, Senior Researcher, CelNet security, India

Dr. Adnan Shahid Khan, University Technology Malaysia, Malaysia

Mr. Prakash Gajanan Burade, Nagpur University/ITM college of engg, Nagpur, India

Dr. Jagdish B.Helonde, Nagpur University/ITM college of engg, Nagpur, India

Professor, Doctor BOUHORMA Mohammed, Univertsity Abdelmalek Essaadi, Morocco

Mr. K. Thirumalaivasan, Pondicherry Engg. College, India

Mr. Umbarkar Anantkumar Janardan, Walchand College of Engineering, India

Mr. Ashish Chaurasia, Gyan Ganga Institute of Technology & Sciences, India

Mr. Sunil Taneja, Kurukshetra University, India

Mr. Fauzi Adi Rafrastara, Dian Nuswantoro University, Indonesia

Dr. Yaduvir Singh, Thapar University, India

Dr. Ioannis V. Koskosas, University of Western Macedonia, Greece

Dr. Vasantha Kalyani David, Avinashilingam University for women, Coimbatore

Dr. Ahmed Mansour Manasrah, Universiti Sains Malaysia, Malaysia

Miss. Nazanin Sadat Kazazi, University Technology Malaysia, Malaysia

Mr. Saeed Rasouli Heikalabad, Islamic Azad University - Tabriz Branch, Iran

Assoc. Prof. Dhirendra Mishra, SVKM's NMIMS University, India

Prof. Shapoor Zarei, UAE Inventors Association, UAE

Prof. B.Raja Sarath Kumar, Lenora College of Engineering, India

Dr. Bashir Alam, Jamia millia Islamia, Delhi, India

Prof. Anant J Umbarkar, Walchand College of Engg., India

Assist. Prof. B. Bharathi, Sathyabama University, India

Dr. Fokrul Alom Mazarbhuiya, King Khalid University, Saudi Arabia

Prof. T.S.Jeyali Laseeth, Anna University of Technology, Tirunelveli, India

Dr. M. Balraju, Jawahar Lal Nehru Technological University Hyderabad, India

Dr. Vijayalakshmi M. N., R.V.College of Engineering, Bangalore

Prof. Walid Moudani, Lebanese University, Lebanon

Dr. Saurabh Pal, VBS Purvanchal University, Jaunpur, India

Associate Prof. Suneet Chaudhary, Dehradun Institute of Technology, India

Associate Prof. Dr. Manuj Darbari, BBD University, India

Ms. Prema Selvaraj, K.S.R College of Arts and Science, India

Assist. Prof. Ms.S.Sasikala, KSR College of Arts & Science, India

Mr. Sukhvinder Singh Deora, NC Institute of Computer Sciences, India

Dr. Abhay Bansal, Amity School of Engineering & Technology, India

Ms. Sumita Mishra, Amity School of Engineering and Technology, India

Professor S. Viswanadha Raju, JNT University Hyderabad, India

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/