Conference Proceedings

The International Conference on Information & Communication Technology

(ICICT'2012): Applications and Techniques

ICICT

Leaders Hall, RamAllah
Palestine

26th July, 2012

Jawwal

كتاب المؤتمر

المؤتمر الدولي لتكنولوجيا المعلومات والإتصالات

التطبيقات والتقنيات (ICICT'2012)

ICICT

قاعة ليدرز / الماصيون
رام الله / فلسطين
الثلاثاء ٢٦، يوليو، ٢٠١٢

جوال

Al-Quds Open University

Deanship of Scientific Research and Graduate Studies

The International Conference on Information & Communication Technology

(ICICT'2012): Applications and Techniques

Al-Quds Open University

Deanship of Scientific Research & Graduate Studies

Palestine- Ramallah\ Al-Masyoun

BBOX. 1804

TEL: 02-2952508 \ 02-2984491

FAX: 02-2984492

E-MAIL: SPRGS@QOU.EDU

2013

Design by:

Information & Communication Technology Center (ICTC)

We sincerely appreciate the support of our sponsors.

Message from Conference Chairman

On behalf of the Conference Organizing Committee, I would like to thank Prof. Dr. Younes Amro (University President) for his support for the International Conference on Information & Communication Technology (ICICT'2012): Applications and Techniques. The conference provided an opportunity for participants to share their ideas and research in the field of Information and Communication Technology.

The Conference which was organized by Al-Quds Open University, Palestine focused on state of the art technologies pertaining to digital information and communication. The current emphasis reflected the growing Interest in new areas in Information and Communication The core theme of the conference has been opened for a wide range of topics of growing interest to researchers in the field of Information and Communication Technology .

We hope you have found the conference productive, informative, and enjoyable. We are looking forward to receiving your constructive comments that would help us in our future planning. Please visit our website (http://www.qou.edu/icict2011/index.jsp) to obtain information of future conferences and journals.

I would like to take this opportunity to thank the organizing team that did an excellent job of putting this conference together. I am also indebted to our reviewers who reviewed the manuscripts, sometimes under extreme time constraints, and selected the best papers that fit this conference.

On behalf of the conference Hosts, Organizers, Scientific Committee and Financial sponsors, we thank all participants and researchers in the International Conference on Information & Communication Technology (ICICT'2012), held at Al-Quds Open University, Ramallah, Palestine on 26th,July,2012. We are confident that this conference provided useful insights and promoted lasting cooperation on these important topics

Yousef Abuzir, PhD

Organizing and Scientific Committee Chairman

Table of Contents

Critical Factors Influencing the Acceptance and Diffusion of E-Government Services: Conceptual Framework

Mohammed Ayoub
Limkokwing University – Malaysia
ayoub22265@hotmail.com

Abstract

There are some great innovations in e-government during the past decade, and there is intense competition between some governments and leaders in the supply of services on the internet. Some countries do not want to stay behind in this area, where many governments have developed detailed strategies to realize the e-government systems, but there is a problem facing these governments which lack user acceptance of e-government services. The purpose of this study is to suggest comprehensive model to explore and investigate the Factors Influencing the Acceptance and Diffusion of e-Government Services. The proposed Model will develop based on the related literature. The motivation for conducting this study, that it is the first study in the Palestine, that investigate users needs and expectations, where there is a significant part of e-government literature that investigates various factors that determine intention to use e-government in developed countries, however, there is a dearth of studies that investigate intention to use e-government in developing countries. Consequently, the final modified research model has the power to explain and predict user behavior in developing countries and especially Palestine. A thorough understanding of the model may help practitioners to analyze the reasons for resistance toward the technology and also help them to take efficient measures to improve user acceptance and usage of the technology.

Keywords: *E-Government, E-Government Acceptance, Intention to Use e-government services, Perceived Usefulness, Perceived Ease of Use, User Satisfaction, And Subjective Norm.*

1. Introduction:

People vary in their orientation towards using technology, some of them reject using technology because they do not see the benefit desired from the use or because they see great complexity within them which causes them a lot of trouble in dealing with them, and others have lack of confidence in it, also some organizations face staff resistance or lack of confidence in new technology or computer systems which affect the investment in technology and prevent or impair performance improvement, and thus the inability to perform their duties daily as required to be collected is the failure of the system. In general, the modernization of public services through the adoption of information and communication technologies is in motion. There are around us, evidences of a universal shift toward modern online public services (e-services) and a dynamic e-business environment. This

has caused governments and public sector organizations around the globe to take care of this phenomenon, become aware of its potentials and consequently utilize them, thereby triggering investments into e-services.

Since the late 1990s, numerous governments have made huge investments in electronic government services to link government networks and deploy a variety of service infrastructure to provide extensive and proactive services. However, low levels of user acceptance of these services are recognized as an endemic problem for government policy makers, government agencies, and e-Government services providers. Behavioral issues of e-Government research are markedly more important than technological ones. More empirical studies on user acceptance of e-Government services are needed to assist governments in improving the effectiveness and quality of e-Government services. Now the need for discovering determinants of adopting e-Government is enormous, but few empirical studies can be found addressing the issue [1]. In order to achieve the needs of all types of users, the designers have to first understand the different requirements that users expect, and then relate these characteristics to the design features. In view of the lack of empirical studies on determinants of users acceptance in relation to e-Government, this study represents an important attempt to address user's attitude towards e-Government services. Davis in 1986 introduced a model to explain user acceptance behavior named the Technology Acceptance Model (TAM). The technology acceptance model (TAM) is one of the most widely used models to explain user acceptance behavior. This model is grounded in social psychology theory in general and the Theory of Reasoned Action (TRA) in particular. Based on the above motivation, this study aims to develop an integrative model of users' acceptance of e-Government services for understanding the factors influencing the acceptance and diffusion of e-government services according to Technology Acceptance Model (TAM).

2. Problem Statement

Rarely governments have walked through users' experience to understand their needs. User experience is a key element of e-government systems design. In addition, there are many of e-governments nowadays, but only a small percentage of them ever reach a high ranking or manage to attract more citizens. One of the important issues facing e-government systems is how to assess and measure the acceptance of e-government based on the experience of citizens with e-government systems. Based on the above, we can illustrate the problem statement as follows:

1- Explaining user acceptance of new technology is often described as one of the most mature research areas in the contemporary information systems (IS) literature [1]. It is noted that there are lack in empirical studies on

determinants of users' acceptance in relation to e-Government.

2. Most previous studies based on the specific number and not enough of the variables that affect the level of acceptance of e-government services, in this study will be extend the technology acceptance model and add some variables that affect the acceptance of e-government services.

3. The technology Acceptance Model is indeed a very popular model for explaining and predicting system use. To date there have been an impressive number of studies on TAM, a great amount of the research has been conducted in the U.S. and only a limited number of studies have focused on the acceptance of technology outside North America [3], but while several confirmatory results have been obtained, there are skepticisms shared among some researches regarding the application and theoretical accuracy of the model. Consequently, it is tempting conduct that research on TAM may have reached a saturation level, for these reasons we will focus in developing and extending model that would exploit the strengths of the TAM model while discarding its weaknesses, Particular, in line with other countries and cultures outside of North America.

4. The external variables that impact the perceived usefulness and perceived ease of use are not completely explored in the TAM. So, in this study will clarify the impact of the e-government information systems quality on perceived usefulness and perceived ease of use.

3. Objectives of Study

1- Due to the current limited number of studies evaluating e-government services acceptance, the researcher wants to set an example for similar research in the future through the understanding of the factors influencing e-government services acceptance. The goal of this study is to introduce the comprehensive model to gain a deep understanding of citizen experience, to identify factors that affect the behavior towards the acceptance of e-government services.

2- The researcher want to give out some feasible suggestions for decision makers and information systems designers to improve the e-government systems and services based on feedback from users when he apply the proposed model.

3- Develop a model of technology acceptance that will have the power to demonstrate acceptance and usage behavior of e-government services by citizens, also understanding of the model may help practitioners to analyze the reasons for resistance toward the technology and would also help to take efficient measures to improve user acceptance/usage of the technology.

4- As governments are increasingly spending large sums of money for delivering e-government services and availability of limited studies on

assessing e-government systems acceptance, developed our interest to conduct research in this area. In addition, to adapting the technology acceptance model (TAM) in the context of e-government systems.

5. Relevance and Significance:

1. Based on a relatively clear description and understanding of models and theories of technology acceptance that has been synthesized from theoretical and practical viewpoints, this study provides a comprehensive model to examine and understand the factors that affect the level of acceptance of e-government services.

2. Knowledge of the needs and expectations of users of e-government services helps systems designers and decision makers to develop and design of systems and services to meet the requirements of the users and raises the level of acceptance of e-services.

3. According to lack in empirical studies on determinants of users acceptance in relation to e-Government, this study provides a theoretical foundation for researching e-Government acceptance continuance in the future.

4. To the practitioners (or governments in this context), this study provides a useful guideline for achieving better e-Government services and increasing the level of acceptance by identifying specific continuance intention factors which are simple, easy to understand, and can be manipulated through system design and implementation. It thereby assists governments in using the findings of proposed model for development and evaluation of e-Government acceptance.

6. Theoretical Background:

Acceptance of a system is a measure of the proclivity to use that system. Without acceptance, there is no inclination to accommodate and include the system within the management process [4]. The successful implementation of information systems (IS) is dependent on the extent to which such a system is used and eventually adapted by potential users [5]. IS implementation is not likely to be considered successful if users are unmotivated to use that type of technology [4]. If users are not willing to accept the information system, it will not bring full benefits to the organization [6], [4]. To predict, explain and increase user acceptance, organizations need to better understand why people accept or reject IS [2]. In this regard, researchers have developed and used various models to understand acceptance of users of IS. Among the different models proposed the Technology Acceptance Model (TAM) [6], adapted from the Theory of Reasoned Action (TRA), and appears to be the most widely accepted among the information system researchers.

The primary goal of TAM is to predict IS acceptance and diagnose design problems before user have experience

with the new system. TAM suggests that when user encounter new IS technologies the two main factors influences how and when they will use the system. These two main constructs of TAM are perceived usefulness and perceived ease of use. TAM proposes that two particular constructs, that are of primary significance for IS acceptance, perceived usefulness (PU) and perceived ease of use (PEOU) affect user's' attitude towards using the information system. While basic constructs of TAM, PU and PEOU, have been considered primary determinants of individual's acceptance and use of technology. IS researchers have investigated and replicated these two constructs and agreed that they are valid in predicting user's acceptance of various IS [5].

In their integration of the technology acceptance literature, the [5] stress the need to extend this literature by explicitly considering system and information characteristics and the way in which they might influence the core beliefs in TAM, and might indirectly shape system usage. Recent studies that have used TAM as a theoretical framework have suggested excluding attitude construct from the TAM model since it does not mediate fully the effect of perceived usefulness and perceived ease of use on behavioral intention as originally anticipated. Recently, the [7] in a research study related to the dimension of IS success suggested that system quality (i.e. information and system quality) affects perceived usefulness, user satisfaction and system usage. According to [8], TAM provides limited guidance about how to influence usage through design and implementation. They further elaborated that as PU and PEOU are abstract concepts and provide general information to the designers. Therefore designers are unable to receive actionable feedback about the important aspects of the IS artifacts itself. They identified information and system quality significant constructs which can affect IS usage. Furthermore, [6] himself noted that future technology acceptance research needs to address how variables affect usefulness, ease of use, and user acceptance.

The [9] examined the adoption of e-government in Australian public citizens based on TAM. Huang, et all's research effort focused on an actual system usage with two constructs, perceived usefulness and perceived ease of use. Their research indicated that the prediction of TAM theory was not supported by the findings. It can be argued that basic constructs of TAM, perceived usefulness and perceived ease of use, may not fully determine users' acceptance of e-government, which therefore brings in the need to search for additional factors that may better predict and enhance the user acceptance of E-Government. Another point that has not been explored well in TAM research is the role of system characteristics as external variables. [10] Did not include other factors explicitly into the TAM model that are expected to impact intentions and usage through PU and PEOU. These external

variables could be system characteristics, organizational structure, training, and the like [10].

7. Overview of E-Government:

Many studies have defined e-government in different ways: [11] has defined e-government as the combination of electronic information-based services (e-administration) with the reinforcement of participatory elements (e-democracy) to achieve the objective of "balanced e-government". The [12] defined e-government as the delivery of government information and services online through the internet or other digital means. E-Government has also been defined as the delivery of improved services to citizens, businesses, and other members of the society through drastically changing the way governments manage information [13]. Cited in [14]. It seems that there are a number of e-government definitions in the existing literature. As is clear, most definitions of e-government revolve around the concepts of government's employment of technology, in particular web-based application to improve the access and delivery of government services to citizens, business partners, and other government agencies. Full utilization of e-government will bring a lot of benefits to the management philosophy of many governments and is going to bridge the interaction gap between ordinary citizens and the government. E-Government can also result in huge cost savings to governments and citizens alike, increase transparency and reduce corrupt activities in public service delivery [14].

8. E-Government Acceptance Theories and Models:

As governments continue to invest heavily in IT, understanding the usage behavior of end users has become an important topic in research on e-government implementation. It is also of increasing practical importance as the usage of IT becomes more pervasive. In recent year, intention-based models, e.g., the theory of reasoned action (TRA) [15], the technology acceptance model (TAM) [6], and the theory of planned behavior (TPB) [16], [17], [18] have been employed to provide an understanding of the determinant of technology usage. Intention-based models use behavioral intention to predict usage and, in turn, focus on the identification of the determinants of intention, such as attitudes, social influences, and facilitating condition [10], [19], [20], [21]. There was considerable empirical support for these intention-based models, and researchers have suggested various ways to broaden their applicability.

8.1. The Theory of Reasoned Action (TRA):

Ajzen and Fishbein developed a versatile behavioral theory and model in 1980 called the Theory of Reasoned Action (TRA). This model forms the backbone of studies associate with attitude-behavior

relationships. This has been adapted for use in many fields and is widely used in academia and business today.

TRA is a social-psychological model that addresses the determinants of consciously intended behavior [15], [22]. This model proposes that individual behavior results from conscious intentions to perform that behavior, that is, a behavioral intention. The behavioral intention arises from the individual's own attitude toward the behavior and his or her perception of important others' normative preferences about engaging in the behavior. This normative influence is referred to as the subjective norm. A central proposition of TRA, which has considerable empirical support [15], is that individual behavior is a direct, positive function of behavioral intention, which in turn, is determined by two conceptually distinct constructs: attitude toward the behavior and subjective norm. The [10] found that behavioral intention to use the system is significantly correlated with usage, and that behavioral intention is a major determinant of user behavior while other factors influence user behavior indirectly through behavioral intention (see Figure 1).

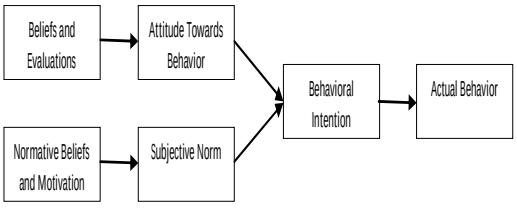


Figure 1: Theory of Reasond Action (TRA) Adopted from [10]

8.2. The Theory of Planned Behavior (TPB):

The Theory of Planned Behavior is proposed as an extension of the Theory of Reasoned Action. The TPB introduced a third independent determinant of intention, perceived behavior control (see figure 2). According to [18], TPB incorporates an additional construct in order to account for situations where an individual lacks the control or resources necessary for carrying out the targeted behavior freely. TPB is a theory that predicts deliberate behavior, because behavior can be deliberative and planned, and TPB is considered to be more general than TRA [23]. It can be noticed that when given a sufficient degree of actual control over their behavior, people are expected to carry out their intentions when the opportunity arises. In addition, according to the TPB, human behavior is guided by three kinds of beliefs:

1. Behavioral beliefs - beliefs about the likely outcomes of the behavior and the evaluations of these outcomes.
2. Normative beliefs refer to the perceived behavioral expectations of such important referent individuals or groups as the person's spouse, family, friends, and co-workers.
3. Control beliefs - beliefs about the presence of factors that may facilitate performance of the behavior and the perceived power of these factors.

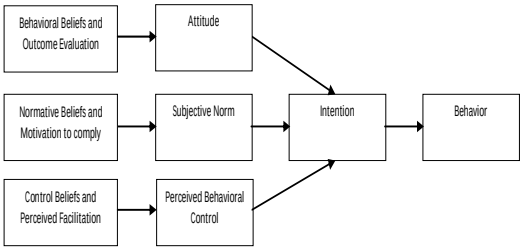


Figure 2: The Theory of Planned Behavior [16]

8.3. The Technology Acceptance Model (TAM):

The technology acceptance model (TAM) is one of the most widely used models to explain user acceptance behavior. This model is grounded in social psychology theory in general and the Theory of Reasoned Action (TRA) in particular [15]. According to the TRA, behavioral intention may be defined as a measure of the strength of one's intention to perform a specific behavior [15]; that is, use an information system.

Davis and his colleagues included two powerful and parsimonious constructs to represent the antecedents of system usage in TAM: Perceived usefulness (defined as the degree to which a person believes that using a particular technology will enhance his or her job performance) and Perceived ease of use (defined as the degree to which a person believes that using a particular technology will be free of effort) [6].

TAM postulated that actual system usage was determined by a behavioral intention to use a system, which was jointly determined by a person's attitude toward using the system and its perceived usefulness. This attitude was also jointly determined by perceived usefulness and perceived ease of use, with perceived ease of use having a direct influence on perceived usefulness. Finally, perceived usefulness and perceived ease of use were directly influenced by the system design characteristics (see Figure 3).

The goal of TAM is to provide an explanation of the determinants of computer acceptance that is in general capable of explaining user behavior across a broad range of end-user computing technologies and user populations, while at the same time being both parsimonious and theoretically justified. But because it incorporates findings accumulated from over a decade of IS research, it may be especially well suited for modeling computer acceptance [10].

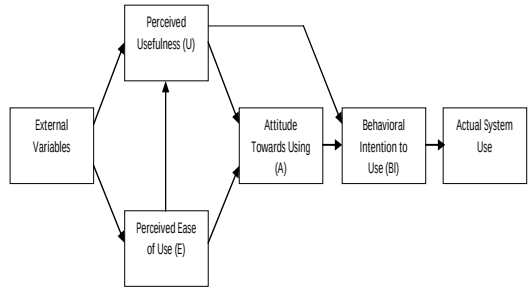


Figure 3: The Technology Acceptance Model [6]

9. Proposed Research Model:

We review the literature in IS fields that is related to e-government acceptance and usage. The aim of the review is to explicate the potential antecedents of citizen's acceptance of e-Government

services and integrate them into a model (see Figure 4).

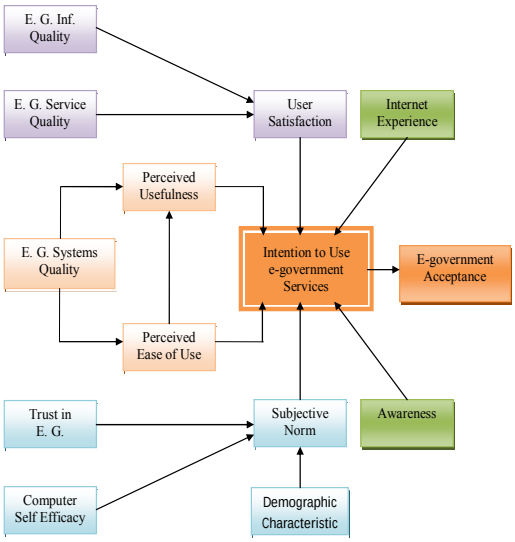


Figure 4: research model

Based on our review of the e-government and IS acceptance literatures related to e-government acceptance and use, we hypothesize that the intention to use e-government services was a major determinant of the e-government acceptance, in turn, we identified perceived usefulness, perceived ease of use, user satisfaction, internet experience, awareness, and subjective norm as antecedents of citizens' intention to use e-Government services, with perceived ease of use having a direct influence on perceived usefulness. The e-government system quality that is likely to influence perceived usefulness and perceived ease of use are also included in the model.

E-Government information quality and E-Government service quality are considered as potential antecedents of user satisfaction. For subjective norm, trust in e-government, computer self efficacy, and demographic characteristics are considered as determinants. Intention to use is taken as the dependent variable of this study.

10. Hypotheses Development:

Based on above, we can develop the research hypothesis as follow:

E-Government acceptance is the actual level of usage by the end users. The acceptance by end users depends primarily on the behavioral intention of the end users. Though the other factors described in the research model also affects the acceptance of the e-government, the direct relation is with the end user behavior intention [24]. Behavioral intention is a measure of the strength of one's intention to perform a specified behavior. Actual use refers to an individual's actual direct usage of the given system. The TAM constructs and the relationships among them are used here because e-government is based on new information technologies, such as the internet and the World Wide Web [24]. Based on previous, we can hypothesize:

H1: Intention to use e-government services has a positive effect on e-government acceptance.

Oliver and Shapiro [25] found that the stronger a person's self-efficacy beliefs, the more likely he or she was try to achieve the desired outcome. In the present context this means that Internet experience should be positively related to the intention to use e-government

services, such as WWW service. Therefore, the following hypothesis is proposed:

H2: Internet experience will have a positively associated with intention to use e-government services.

Awareness is a variable associated with people's knowledge about e-Government and the availability of electronic services online. Recent research conducted in Lebanon, which is a Middle East country with a similar profile to Jordan, indicted that awareness of the existence of e-Government services is positively related to the usage of e-Government services [26]. Based on previous, we can hypothesize:

H3: Awareness will have a positively associated with intention to use e-government services.

User satisfaction is an important component to measure IS success. It can be defined as the extent to which users believe that the IS available to them meets their information requirements [27]. According to TRA/TPB, attitude and behavioral belief measures should be specified in a way that corresponds to the time, target, and context of the behavior of interest, in order to be a good predictor of the behavior or behavioral intention [28]. In our study user satisfaction is treated as an attitude toward intention to use e-government services Therefore, user satisfaction in this study represents both the evaluation of the IS and the evaluation of the usage experience with the IS. Based on the abovementioned discussion we hypothesize that:

H4: User satisfaction has a positive effect on intention to use e-government services.

The Theory of Planned Behavior (TPB)asserts that behavior is a direct function of behavioral intention, and behavioral intention is determined by the individual's attitudes towards performing the behavior, the subjective norms held by the individual, and the individual's perceived behavioral control over the act [18] cited in [29]. Attitude refers to an individual's positive or negative feelings about performing the target behavior. TPB predicts that the more favorable an individual evaluates a particular behavior, the more likely he or she will intend to perform that behavior [30]. Subjective norms reflect the person's perception that most people who are important to him think he should or should not perform the behavior in question. Several empirical studies have shown that subjective norms have a positive and direct impact on behavioral intention, but this influence is usually weaker than that of attitude and perceived behavioral control [22], [31]. Based on previous, we can hypothesize:

H5: Citizen’s Subjective norm has a positive effect on intention to use e-government services.

Information quality is related to the quality of information that the e-government delivers to its users [32],

their model proposes that "system quality and information quality singularly and jointly affect both use and user satisfaction [33]. Therefore, e-government can be viewed as information systems. Previous studies used information quality to measure IS success [34], measuring e-commerce success [32], and e-shopping acceptance [35]. This study adopts the following hypothesize:

H6: E-Government Information quality is significantly associated with user satisfaction.

Service quality is one of the focuses in IS research as well as e-government research recently. In e-government context, service quality might be an important factor to explain citizen's acceptance of e-services. Hence, examining the quality of e-service could determine whether users tend to continue to use the system or not. It is believed also by offering the best service will entice the citizens to use online services and gain the advantages from it [36]. These arguments have led us to postulate:

H7: E-Government service quality will positively influence user satisfaction.

Davis [6] developed and validated better measures for predicting and explaining use which focused on two theoretical constructs: perceived usefulness and perceived ease of use, which were theorized to be fundamental determinants of system use. TAM theorized that the effects of external variables (e.g., system characteristics, development process, training) on intention to use are mediated by perceived usefulness and perceive ease of use. Perceived usefulness is also influenced by perceived ease of use because if other things are equal, the easier the system (technology) is, the more useful it can be [4]. Based on above, we can postulate:

H8: Perceived usefulness of e-Government services will positively influence intention to use e-government services.

H9a: Perceived ease of use of e-Government services will positively influence Perceived usefulness of e-Government services.

H9b: Perceived ease of use of e-Government services will positively influence intention to use e-government services.

Trust is an important element of e-government [37]. The [37] defines trust as a belief that others will behave in a predictable manner. The importance of trust in e-government adoption has been stated by many researchers [38], [39]. The [38] argue that trustworthiness is one of the main factors that influence citizens' intention to use e-government service in addition to perceived ease of use and compatibility. Therefore, the effects of trust in e-government on intention to use are mediated by subjective norm, in turn, we can hypothesize:

H10: Citizen's trust in e-government will positively influence citizen's subjective norm.

Compeau and Higgins [40] defined computer self-efficacy as "an individual's perceptions of his or her ability to use computers in the accomplishment of a task". Individuals with a high computer self-efficacy magnitude would see themselves as able to accomplish difficult computing tasks and would judge themselves as capable of operating with less support and assistance than those with lower computer self-efficacy magnitude. Compeau and Higgins [40] also reported that computer self-efficacy plays an important role in shaping an individual's feeling and behavior. Importantly, in the context of e-Government, Wangpipatwong et al. [41] empirically confirmed that the adoption of e-Government websites depends on the computer self-efficacy of citizens. Cited in [42]. Therefore, the effects of computer self-efficacy on intention to use are mediated by subjective norm. Thus, this study proposes:

H11: Computer self-efficacy of citizen will positively influence citizen's subjective norm.

Demography is the available information on any given user or group. Demographic data refers to selected population characteristics which are used to classify people for statistical purposes, such as age, gender, education and experience. Prior research on e-government has identified general demographic characteristics of citizens who use e-government services. Dimitrova & Chen [43] in their exploratory study proposed a multidimensional theoretical framework combining diffusion of innovations and the technology adoption model to explain e-government adoption in the United States. A number of determinants were proposed and tested, going beyond the traditional demographic profiling of e-government users. The main conclusion is that sociopsychological factors affect e-government adoption. Choudrie & Dwivedi [44] in their study found that the demographic characteristics of citizens such as the age, gender, education, and social class have an imperative role in explaining the citizen's awareness and adoption of e-government services in the household [45]. Therefore, the effects of demographic characteristics on intention to use are mediated by subjective norm. Thus, this study proposes:

H12: Citizen's demographic characteristic significantly associated with citizen's subjective norm.

System quality refers to the technical details of the information system interface and quality of system that produces information output [33]. Davis [6] did not include system characteristics into TAM model, but he suggested including judicious system characteristics. According to DeLone and McLean [33] technology

characteristics singularly or jointly affect subsequent use and user satisfaction. Hence, it is assumed that e-government systems quality influence on PU and PEOU. Thus, this study postulates the following hypotheses:

H13a: E-Government Systems quality will have positive effect on PU of e-government services.

H13b: E-Government Systems quality will have positive effect on PEOU of e-government services.

11. Conclusions:

The propositions presented in this paper an opportunity for further investigation in the factors influencing the acceptance and diffusion of e-government services, and e-government acceptance from citizens. The proposed model should be of interest to information systems practitioners, academic community, and decision-makers in e-government. For the practitioner community, the model will enhance their understandings on the factors that contribute towards e-government acceptance. For the academic community, the proposed model provides ample research opportunity to validate in order to support or refute the proposed propositions. And decision-makers in the e-government, that the proposed model will help them to increase level of e-government services acceptance by understanding the needs of citizens and their expectations.

We recommend that other researchers for the work of empirical study on the proposed model in developing countries, especially in Palestine, to know and anticipate the factors that affect the acceptance of e-government services by citizens especially that these countries in initiative stage of construction and implementation of e-government.

Acknowledgment

I want to thank staff and all colleagues in Al-Quds Open University for giving us the opportunity to participate in this conference, we wish to be the first of other conferences in order to catch up with technological progress and support of the innovation, and thanks to Dr. Mohammed Abu Gabeen for continuance with me throughout my work in this research. I also thank colleagues at the Limkokwing University in Malaysia for their help and advice, especially Prof. Dr. Ahmad Faisal, my supervisor Dr. Kulandayan Ramanathan, and co- supervisor Dr. Ilham Sentosa

References

[1] N. M. Yaghoubi, B. Kord & R. Shakeri, *E-Government Services and user Acceptance*: The Unified Models' Perspective. European Journal of Economics, Finance and Administrative Sciences ISSN 1450-2275 Issue 24, 2010

[2] F. D. Davis, *A Technology Acceptance Model for empirically testing new end-user information systems*: theory and results. Doctoral Dissertation Thesis, 1986

[3] D. F. McCoy & Everard, *The effect of culture on IT diffusion*: using the Technology Acceptance Model to predict email usage in Latin America, Americas Conference on Information Systems, pp. 1899-901, 2000

[4] V. Venkatesh, & F. D. Davis, *a Theoretical Extension of the Technology Acceptance Model*: Four Longitudinal Field Studies. Management Science Vol. 46, No. 2, pp. 186-204, 2000

[5] V. Venkatesh, M. Morris, G. Davis, & F. Davis, *User acceptance of information technology*: Toward a unified view. MIS Quarterly, 27(3), 425-478, 2003

[6] F. D. Davis, *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology*, MIS Quarterly, 13 (3), 319-40, 1989

[7] W. H. DeLone, E. R. McLean, *Measuring e-Commerce Success*: Applying the DeLone & McLean Information Systems Success Model, International Journal of Electronic Commerce, Vol. 9, No. 1, pp. 31-47, 2004

[8] B.H. Wixom, P.A. Todd, *A theoretical integration of user satisfaction and technology acceptance*, Information Systems Research 16 (1) 85–102, 2005

[9] W. Huang, J. D'Ambra, & V. Bhalla, *An empirical investigation of the adoption of e-government in Australian citizens*: Some unexpected research findings. Journal of Computer Information Systems, 43(1), 15-22, 2002

[10] F.D. Davis, R.P. Bagozzi, and P.R. Warshaw, *User acceptance of computer technology*: A comparison of two theoretical models, Management Science, vol. 35, no.8, pp. 982-1003, 1989.

[11] S. Coleman, *African e-Governance – Opportunities and Challenges*, University of Oxford, Oxford University Press, 2006

[12] A. Muir, and C. Oppenheim, *National Information Policy Developments Worldwide in Electronic Government*, Journal of Information Science, 28, 3, 173 – 186, 2002

[13] V. Kumar, B. Mukerji, B. Irfan, and P. Ajax, *Factors for Successful e-Government Adoption*: A Conceptual Framework, The Electronic Journal of e-Government, 5, 1, 63-77, 2007

[14] K. J. Bawalya, *Factors Affecting Adoption of E-Government in Zambia*, EJISDC, The Electronic Journal on Information Systems in Developing Countries 38, 4, 1-13, Gaborone, Botswana, 2009

[15] M. Fishbein & I. Ajzen, *Belief, attitude, intentions and behavior*: An introduction to theory and research. Boston: Adison-Wesley, 1975

[16] I. Ajzen, *From intention to action*: A theory of planned behavior. In J. Kuhl and J. Backman (Eds), Action control: From cognition to behavior, New York: Springer-Verlag, 1985

[17] I. Ajzen, *Attitude, Personality, and behavioral*. Milton Keynes, UK: Open University Press, 1989

[18] I. Ajzen, *the theory of planned behavior*. Organizational Behavior and Human Decision Proccesses, 50, 179-211, 1991

[19] F. D. Davis, R. P. Bogozzi & P. R. Warshaw, *Extrinsic and intrinsic motivation to use computers in the workplace*, Journal of Applied Social Psychology, 22, 1111-1132, 1992

[20] J. Hartwick, & H. Barki, *Exploring the role of user participation in information systems use*. Management Science, 40(4), 440-465, 1994

[21] K. Mathieson, *Predicting user intentions*: Comparing the technology acceptance model with the theory of planned behavior. Information systems Research, 2, 173-191, 1991

[22] M. Fishbein & I. Ajzen, *Understanding Attitude and Predicting Social Behavior*, Pretice-Hall, New Jersey, 1980

[23] Chau, PYK, Cole, M, Massey, AP, Montoya-Weiss, M & O'Keefe, R *Cultural differences in the online behaviour of consumers*, Communications of the ACM, vol. 45, no. 10, pp. 138-43, 2002

[24] G. P. Sahu, M. P. Gupta & T. Sahoo, *Towards a Model of e-Governance Acceptance*, the second International Conference on e-Governance, Colombia, Sri Lanka, 2004

[25] T.A. Oliver, F. Shapiro, *Self-efficacy and computers*, Journal of Computer-Based Instruction 20, 81–85, 1993

[26] A. Charbaji, and T. Mikdashi, *a Paralytic Study of the Attitude toward e-Government in Lebanon*, Corporate Governance. v.3, n.1, p.76-82, 2003

[27] P. Bharati, *People and information matter*: task support satisfaction from the other side, The Journal of Computer Information Systems 43 (2) 2003

[28] J. Lee, N. Bharosa, J. Yang, M. Janssen & H. R. Rao, *Group value and intention to use* — A study of multi-agency disaster management information systems for public safety, Elsevier, 2010

[29] C. K. Farn, Y. W. Fan & C. D. Chen, *the study of electronic toll collection service adoption*: an integrated model, Taiwan (R.O.C.), 2001

[30] I. Ajzen, *Attitude, traits, and actions*: Dispositional prediction of behavior in social psychology. Advances in Experimental Social Psychology, 20: 1-63, 1987

[31] C. J. Armitage, M. Conner, *Efficacy of the theory of planned behavior*: a meta-analytic review, Br. J. Social Psychol., 40: 471-499, 2001

[32] W. H. DeLone, E. R. McLean, *Measuring e-Commerce Success*: Applying the DeLone & McLean Information Systems Success Model", International Journal of Electronic Commerce, Vol. 9, No. 1, pp. 31-47, 2004

[33] W. H. Delone, and E. R. Mclean, *Information Systems Success*: The Quest for the Dependent Variable," Information Systems Research, 3 (1), 60-93, 1992

[34] J. Iivari, *An Empirical test of the DeLone-McLean Model of Information System Success*, The DATA BASE of Advances in Information Systems, Vol. 36, No. 2, pp. 8-27, 2005

[35] H. Shih, *An empirical study on predicting user acceptance of e-shopping on the web*, Information & Management, Vol. 41, pp. 351-368, 2003

[36] R. Hussein, N. Mohamed, A. R. Ahlan, M. Mahmud & U. Aditiawarman, *an Integrated Model on Online Tax Adoption in Malaysia*, European, Mediterranean & Middle Eastern Conference on Information Systems, 2010

[37] M. Warkentin et al, *Encouraging Citizen Adoption of e-Government by Building Trust*. Electronic Markets, Vol. 12, No.3, pp.157 – 162, 2002

[38] L. Carter, and F. Bélanger, *The utilization of e-government services*: citizen trust, innovation and acceptance factors. Information Systems Journal, Vol.15, No.1, pp. 5–25, 2005

[39] E. W. Welch et al, Linking *Citizen Satisfaction with E-Government and Trust in Government,* Journal of Public Administration Research and Theory, Vol. 15, No. 3, pp. 371-391, 2005

[40] D. R. Compeau, and C. A. Higgins, *Computer Self-efficacy*: Development of a Measure and Initial Test", MIS Quarterly, Vol. 19, No. 2, pp. 189−211, 1995

[41] S. Wangpipatwong, W. Chutimaskul, and B. Papasratorn, *A Pilot Study of Factors Affecting the Adoption of Thai e-Government Websites*, in Proceedings of the International Workshop on Applied Information Technology (IAIT'05), pp. 15−21, 2005

[42] S. Wangpipatwong, W. Chutimaskul & B. Papasratorn, *Understanding Citizen's Continuance Intention to Use e-Government Website*: a Composite View of Technology Acceptance Model and Computer Self-Efficacy, The Electronic Journal of e-Government Volume 6 Issue 1, pp 55 – 64 2008

[43] D. V. Dimitrova, Y. C. Chen, *Profiling the adopters of e-government information and services*: the influence of psychological characteristics, civic mindedness, and information channels", Social Science Computer Review, (24:2), pp.172-188, 2006

[44] J. Choudrie, Y. K. Dwivedi, *A Survey of Citizens' Awareness and Adoption of E-Government Initiatives, the 'Government Gateway'*: A United Kingdom Perspective", In the proceedings of the E-Gov. workshop, Brunel University, London, UK, 2005

[45] S. E. Colesca & D. Liliana, *E-government Adoption in Romania*,World Academy of Science, Engineering and Technology 42, 2008

Effects of integrating Web 2.0 applications in the E-Business course

Nadira Alaraj
Bethlehem University, Palestine

Abstract

Learning styles of undergraduate student of today are changing rapidly because of ready access to the Internet. Students are no longer dependent on conventional textbooks to gain knowledge; therefore, traditional teaching methods must be tweaked to accommodate such changes. The E-Business course offered at Bethlehem University in Palestine during the spring 2011 semester blended Web2.0 applications such as social media and cloud computing to support the changing student learning styles, with the intent of effectively delivering the E-Business course content without requiring textbook. The course also included participation in the Google Online Marketing Challenge (GOMC) that is offered worldwide. This article evaluates innovations in terms of Web2 applications skills development of the 59 students taking this course and how skills acquisition influenced their independence in knowledge seeking. A questionnaire assessed the students' perception and was supplemented by the students' results in GOMC. Results of survey showed that 79.30% agreed that the course did help them develop confidence on their knowledge acquiring abilities. This innovation appears to be effective in motivating students to learn how to utilize new applications on their own. These results should encourage educators to employ Web2 applications relevant to the course content and evaluate their results.

Keywords*: Web2.0 applications, blended learning, Web-based learning, Adwords, GOMC, social networks*

Introduction

The term Web2 was introduced by O'Reilly Media Inc. in 2004. Although Web2 has several meanings, in this article it refers to those applications that run on Internet browsers from any computer or mobile device that can access the Internet. The user's data for these applications can be kept on either host's computer servers or on the user's own storage device. Most Web2 applications (Web2-Apps) facilitate online collaborative environment.

"Web2-Apps are changing the Web from an essentially ''broadcast'' environment (where a relatively small number publish material to the rest), to one in which we can all participate as publishers. Arguably Web 2.0 has created a new virtual environment in which young people live and, potentially, learn and take control of the publishing process" (Brown, 2010, p.5). Through social media applications under

the Web2.0 environment, users are sharing contents with their followers and getting personal comments and different opinions. "This raises important issues about traditional learner–teacher relationships, ownership of lecture content, and of control over the dialogue in a classroom" (Brown, 2010, p.5).

In the last decade, the e-learning trend was (and still is) dependent on Course Management System Platforms. These platforms as well as WIKI management and chat rooms were used for communication and collaboration among learners, educators and course owners or served mainly as sources of the course material. This model of e-learning is nothing more than transforming traditional classroom to virtual environment as it still keeps the upper hand of the educator in gearing the learning process.

"So, hype notwithstanding, Web 2.0 tools might turn out to be a lot more popular among learners and teachers because they meet user needs" (Brown, 2010, p.6). In Web2 learners can find same content presented in different medium such as videos, animated graphics or simple textual material and it is also aggregated in different Web2.0 tools. Hence, Web 2.0 brings a new perspective to finding online information. Instead of using commercial search engines like Google and Yahoo to locate information, it is now possible to find information in more social and participatory ways (Asselin & Moayeri, 2011). No matter what the learner uses to search for information, the Web exploration strategies reflect upon and integrate diverse web search results and ideas for a deeper understanding of the given problems (Liu et al, 2010).

Cognitive learning as Cifuentes et al. (2011) point out is concerned with the acquisition of knowledge and it relates to the process of acquiring knowledge by the use of reasoning, intuition or perception. While Web 2.0 technologies interfered with deep learning they may have contributed to learners' cognitive flexibility. The cognitive flexibility emphasizes the importance of providing learners with multiple representations of content in order to increase those learners' cognitive structures. Therefore, educators have to be careful when integrating Web2-Apps in their courses because of the different skills of the students, and consider the benefit from more training in how to use Web 2.0 tools to support their own learning. Such training would decrease students' experience of cognitive overload in courses that leverage Web 2.0 while increasing their cognitive flexibility.

At Bethlehem University the typical use of Web2-Apps has been through the online Open Source Course Management Platform (Moodle) provided by the University's network. Moodle is widely used at Bethlehem University as a platform for the communication and delivery of course material between faculty and students.

Moodle was utilized in the E-Business course along with the Web2-Apps in order to introduce something new to the ICT educational model models at the University.

The main reason for introducing Web2-Apps was to equip the students with new skills and allow them to experience the benefits and limitations of these applications. Another reason for introducing Web2-Apps in the E-Business course was to allow independence from the applications accessed from the University's network and to benefit from e-services provided on the World Wide Web. The skills gained from this will allow the students to be active e-agents when they become part of the local Palestinian or international workforce. The integration of Web2-Apps reflected in the assessment criteria for this course and it had 45% of the total 100% mark. The remaining 55% covered the assessments of exam, written reports and presentation. No traditional textbook was required for this course but online references were provided, along with the assessment criteria for each course activity.

The delivery of the E-Business course was conducted by blending Web2.0-Apps remotely and physically meeting students in classroom. The students were given enough time to present their findings and benefit from the course content. This study enabled them to explore the impact of Web2.0-Apps on their knowledge acquiring skills and other skills necessary for their professional development.

Implementation of Web2-Apps in the E-Business course

Micro-blogging through Twitter was used in the course. Students were requested to tweet in this course for issues relevant to E-Business course topics. The hashtag #BU266 was introduced to aggregate tweets or other web postings in real-time order. As an example, "*How to Develop a Successful Online Marketing campaign http://t.co/bmwGZkk #BU266*" reflects sharing an article relevant to a particular time during the course. In this course the hashtag #BU266 was assigned. The application which was used during the course to aggregate the students' tweets was http://wthashtag.com/ but this company was sold to http://whatthetrend.com/ during the offering of the course. The experience of this transition during the course exposed the students to the dynamics of web companies and changes of the applications service during the transition of ownership. In addition, a common blog http://bu266.blogspot.com/ was employed for this course to collect the students' posts of their essays and these were tagged with relevant keywords. The privacy setting of this common blog was kept public with the intention of encouraging a web presence for the students' work. Students of this course were encouraged to integrate videos of at most 3 minutes duration into their 10 minutes presentations to illustrate the concept under discussions.

With Web2-Apps, real time feeds is very important concept and in order to enable the students understand 'by example' the way Web2-Apps links real-time feeds, they were encouraged to open an account on LinkedIn and link their Twitter account with LinkedIn. In this way the student tweet is automatically posted on his/her LinkedIn home profile.

Another emphasis in this course is the enrollment in the Google Online Marketing Challenge. Students were divided into eleven teams. Each team worked with local client to promote the client webpage within 3 weeks of online marketing campaigns using Google Adwords application. Google credit $200 into each team's account on Adwords to accomplish their pre-planned campaign objective. The WIKI from Moodle platform was used in this course to facilitate the communications and collaborative team writing reports of their marketing campaigns. Google calendar were used to track and remind teams of the due dates of the different stages of the execution of the online marketing campaign.

Methodology

Analyses are based on qualitative as well as quantitative data collected during the course. The Web2-Apps study was administered on 59 students majoring in Business with a minor in Marketing. The E-Business course required for the minor was offered in two sections in the spring 2011 semester. The academic level of those students was average according to their grade point average records of Bethlehem University grading system. They were distributed between 40 female and 19 male students from the junior (49) and senior (10) years.

Three different data collection tools were used in this report the learning experience questionnaire, Google Marketing challenge results and Spreadsheet log file on Google documents.

Learning Experience questionnaire

The questionnaire was administered at the end of the semester when 58 students out of 59 submitted their responses. Questions assessing the students' perception on the Web2-Apps development skills involved a 6-point Likert scale and were based on an extension of Miller's Pyramid (training.net 2010) assessment of skills and performance. On the Web2-Apps development skills portion of the questionnaire, the students were asked about their skills in each application before and after with a 6-point Likert scale reply items starts with *didn't know, heard of, know about, knows how, show how* and the last choice is *does*.

Responses to questions concerning student's perception on the content of the course-topics employed a 4-point Likert scale: *didn't understand, beginning, developing* and ends with *accomplished*.

The questionnaire also included general questions about personal learning skills and open-ended questions. For the open-ended questions the respondents were limited to 140 characters in an attempt to have some conformity with their tweets style. The SPSS software was used to analyze the data from the Learning Experience questionnaire. The post course responses on the Web2-Apps development skills demonstrated a statically significant improvement over the pre course (mean difference = -22.85, t = -26.312, df = 53, $p < 0.001$). The correlation between pre- and post- course was moderate and statistically significant ($r = .54$, $p<0.001$).

Google Online Marketing Challenge Results

Students taking this course were grouped into 11 teams of 5-6 members each competed in Google Online Marketing Challenge. The Google challenge judgment had two components. The first component was a campaign statistics algorithm developed by Google that examines the team's Adwords account activities. The second component was the assessment of the two written reports by global academic judging panel of 17 members.

Spreadsheet log file on Google Documents

A Google spreadsheet file was shared online with the students for them to keep track of each account name for each student on every new Web2-App they had used during the course. There was no restriction on accessibility of this file.

Results

Pre-post course differences in mean scores for each Web2-Apps skill are provided in Table 1. The Web2-Apps used in the course are ranked according to the highest difference in the mean scores.

Table 1: Web2-Apps skills acquired from highest to lowest

Ranking	Web2-Apps	Difference of Mean Scores
1	Google Adwords	3.10
2	Twitter	3.10
3	LinkedIn	2.84
4	Hashtags	2.81
5	Google Calendar	2.74
6	Blogs	2.68

7	Google Document	2.38
8	Wiki	2.37
9	YouTube	0.89

Results show that students reported that Google Adwords and Twitter were the skills acquired most often. Further cross-tabulation of the descriptive statistic analysis for the pre- and post- course observations for Adwords and Twitter skills can be found in Table2 and Table 3. Results in both tables indicate that prior to the course, of the 58 students responding, 89.7% reported complete unfamiliarity with Google Adwords while 34.5% reported no knowledge of Twitter. However, after the course 72.4% noted that they understood how to use Google Adwords while 91.4% stated they got to be Twitter users.

Table 2: Percentage of total for Google Adwords students' skills pre vs. post course

	Google Adwords Skills after taking the course					
Google Adwords before taking the course	Heard of	Know about	Knows how	Shows how	Does	**Total**
Didn't Know…		20.7%	31.0%	25.9%	12.1%	89.7%
Heard of…	1.7%	3.4%		1.7%		6.9%
Know about…		1.7%		1.7%		3.4%
Total	1.7%	25.9%	31.0%	29.3%	12.1%	100.0%

That the assessments of Twitter and Adwords activities counted towards students' total grade could be one of the reasons for such an improvement. LinkedIn use in this course was not part of the course grade, but Table 1 shows LinkedIn ranked the 3rd in terms of student's acquisition between the apps employed in the course. This shows that there are almost assured by other motives than the grades that drive student's interest to learn new applications. Although 77.2% of the students reported that they didn't know anything about LinkedIn before the course. This finding shows the students' awareness of the likely importance of LinkedIn apps in their future good could be their motive to venture on their own in the LinkedIn application.

YouTube is at the bottom of the list in Table1 perhaps reflecting that students were already heavy users of YouTube, which is why it shows the little improvement among the students in acquiring more in YouTube.

Use of the spreadsheet log file during the semester to document the student's activities with the different applications, gave a space of comparison between the

students themselves. This might stimulated some students to follow their classmates. The spreadsheet log file revealed another interest, as 35% of the 59 reported that they created their personal blog, in addition to the common course blog which was under use.

Table 3: Percentage of total for Twitter's students' skill pre vs. post course

	Twitter skills after taking the course				
Twitter before taking the course	Know about	Knows how	Shows how	Does	**Total**
Didn't Know…	3.4%	8.6%	13.8%	8.6%	34.5%
Heard of…	5.2%	19.0%	10.3%	20.7%	55.2%
Know about…		1.7%	3.4%	3.4%	8.6%
Knows how…				1.7%	1.7%
Total	8.6%	29.3%	27.6%	34.5%	100.0%

Other incentives might drive students to learn more and utilized their skills in the applications that they are using when this application relate to real life experience. The Google AdWords experience of the global Google Online Marketing Challenge gave a tangible outcome of this E-Business course. One team from this course won the Middle East/Africa regional winner of the 2011 Google Online Marketing Challenge among 4,429 teams competed from 68 countries (Schwartz 2011). Another team is listed among the semi-finalist and no other teams dropped out or were classified as ineligible campaign according to Google campaign assessment statistical algorithm (Google 2011).

Table 4: Percentage of total for LinkedIn's students' skill pre vs. post course

	LinkedIn skills after taking the course					
LinkedIn before taking the course	Heard of	Know about	Knows how	Shows how	Does	**Total**
Didn't Know…	5.3%	24.6%	17.5%	17.5%	12.3%	77.2%
Heard of…		3.5%	10.5%	1.8%	1.8%	17.5%
Know about…				1.8%		1.8%
Knows how…				1.8%	1.8%	3.5%
Total	5.3%	28.1%	28.1%	22.8%	15.8%	100.0%

On the indirect effect of the course Web2-Apps activities and assignments on the student's professional development skills, the students' response to the nominal questions of *yes, no, no effect* options were as follows.
87.72% agreed that their online-searching abilities improved

72.41% agreed that their writing skills improved
70.70% agreed that their team working skills improved
79.30% developed confidence on their knowledge acquiring abilities.

However, 82.2% found that it was necessary to get guidance from the mentor at the beginning of their work. From the open question, 41% of the students expressed differently but with the same meaning that they were overwhelmed by the load of work in this course as well as the many Web2-Apps given to them in one course.

Discussion and Conclusion

Assessment of students taking the E-Business course demonstrated the effectiveness of utilizing Web2-Apps in enhancing professional skills such as improved writing ability as by-product to the learning process toward achieving the knowledge construction through the interaction with the applications. The Web2-Apps facilitated communication on the subject matter among the students themselves and increase their interaction with the outside world locally as well as internationally on the course topics. However, the dependence on same Web2-Apps or repeating the same activities for the following academic year might not be possible due to the dynamic changes on the net. What is available for free access might not remain so. The applications available now could be replaced with something different in few months and opportunities of today like Google Online Marketing Challenge might not be available for next month. Such contingencies must be considered in the offering of the same course for the next academic year, as well as new opportunities that might emerge by the continuous improvement in the technology.

The Web2-Apps pedagogical approach of this course is different than the traditional one in not making use of the traditional textbook and in the educator role of directing the learners toward student-centered approach. The textbook was replaced by online resources that provided students with materials that recently became available. It is suggested to conduct a useful comparative study where one group will follow the traditional educational method with text book against another group using Web2-Apps approach to validate its' effectiveness on the course topics. Further studies and piloting with varieties of Web2-Apps might consolidate the trend toward modernizing the educational system.

Opportunities for engaging the current generation of students in using the Internet and its application is tremendous, regardless of whether of students' are up to the level of ability. Internet applications provide educator with valuable opportunities to widen their scope of course delivery and customize the learning process to accommodate the different level of learning among the students to help them gain control on their own learning. Yet educators to follow such innovative approach is

quite challenging for it requires the educators to be ahead of their students in keeping up-to-date with the development in Web applications and not just Web2-Apps. One is reminded of Margaret Mead's (1975) observation that in traditional societies, the old teach the young, but in rapidly changing societies, the old learn from the younger generation. The move towards Web applications approach and blend that in the education system requires the support of the universities management to get the change that makes impact on the learners.

References

Asselin, M., & Moayeri, M. (2011) The Participatory Classroom: Web 2.0 in the Classroom. *Practical Strategies*, 13 (2).

Brown, S. (2010) From VLEs To Learning Webs: The implications of Web 2.0 for Learning and Teaching. *Interactive Learning Environments*, 18 (1), 1–10.

Cifuentes, L., Xochihua, O., & Edwards, J. (2011). Learning In Web 2.0 Environments Surface Learning and Chaos or Deep Learning and Self-Regulation? *The Quarterly Review of Distance Education*, 12(1), 1–21.

Google (2011). *Google Online Marketing Challenge,* Retrieved July 29, 2011 from http://www.google.com/onlinechallenge.

Google (2011). *Online Marketing Challenge, 2011 Winners,* Retrieved August 30, 2011 from http://www.google.com/onlinechallenge/winners.html.

Gp-training.net (2010). *The Miller Pyramid and Prism,* Retrieved April 25, 2011 from http://www.gp-training.net/training/educational_theory/adult_learning/miller.htm.

Learning: Supporting Web Co-Discovery in One-to-One Environments. *Educational Technology & Society*, *13* (4), 126–139.

Liu, C.-C., Don, P.-H., Chung, C.-W., Lin, S.-J., Chen, G.-D., & Liu, B.-J. (2010). Contributing, Exchanging and Linking for Learning: Supporting Web Co-Discovery in One-to-One Environments. *Educational Technology & Society, 13*(4), 126-139. (SSCI).

Miller, S. (2011). 50 Ways to Use Twitter in the Classroom. *TeachHUB,* [blog] 15th October 2005, Retrieved July 29, 2011 from http://www.universityreviewsonline.com/2005/10/50-ways-to-use-twitter-in-the-classroom.html.

O'Reilly (2011). *What Is Web 2.0*. Retrieved July 29, 2011 from http://oreilly.com/web2/archive/what-is-web-20.html.

PCMag.com (2011). *Definition of: LinkedIn, Retrieved* July 19, 2011 from http://www.pcmag.com/encyclopedia_term/0,2542,t=LinkedIn&i=60336,00.asp.

PCMag.com (2011). *Definition of: YouTube,* Retrieved July 19, 2011 from http://www.pcmag.com/encyclopedia_term/0,2542,t=YouTube&i=57119,00.asp.

Schwartz, J. (2011). Announcing the Winners of The Google Online Marketing Challenge. *Google for Nonprofits Blog,* [blog] Retrieved August 22, 2011 from http://googlefornonprofits.blogspot.com/2011/08/announcing-winners-of-google-online.html

TechTarget (2000). *Definition Weblog,* Retrieved July 14, 2011 from http://searchsoa.techtarget.com/definition/weblog.

Twitter (2011). *Twitter Basics,* Retrieved July 14, 2011 from http://support.twitter.com/groups/31-twitter-basics#topic_104.

Yang, D, Richardson, J, French, B, & Lehman, J 2011, 'The Development of a Content Analysis Model for Assessing Students' Cognitive Learning in Asynchronous Online Discussions', *Educational Technology Research And Development*, 59, 1, pp. 43-70.

Protection the Copyright in E-Education Process

Dr. Osama Amin Marie
AL – Quds Open University
e-mail: omarie@qou.edu

Dr. Khader Muspah Titi
King Khaled University
e-mail: drkhadermuspah@gmail.com

Abstract

Today's world, becoming more competitive, every day is demanding from organization the flexibility to adapt themselves to the permanent situations of market change, readiness for ongoing development and guarantee of the quality of products and services. At the same time, Internet, after being used initially as a great source of information exchange, rapidly happen to be used as an important means for providing learning and training services across the whole world. However, such advances have caused series of information system security issues to the face. The complexity of Internet infrastructures, such as in a Web services distributed system, can hide the potential risks of so many security issues, and subsequently become disadvantageous to e-learning users, applications and institutions.

The main aim of this research is to provide an approach to protect the copyright of e-courses materials in the e-learning system. This new model will be deployed to protect the copyright of e-courses material from unauthorized distribution, and to protect the e-course material from being modified while transit. The design of model is provided to make the e-learning process more secure for both organization and students alike.

Key words: *security, e-learning, encryption, RSA*

1. Background to the Study

Information is now the most valuable resource in the world. Whether it is a personal letter, documents or an industrial secret, all information has a worth to somebody. This research considers issues of security and privacy for such information.

The world in which we exist is decreasing because virtually every person can be electronically connected either by satellite communication, Internet, electronic mail (e-mail), or a conventional telephone in a global village network that transcends geographic boundaries. Progress in communication and information technologies (CIT) has brought extraordinary changes to the whole of our world, which transforms us toward an information civilization. As we shift into the twenty one century, we discover that our dependence on

information technology (IT) is increasing dramatically. IT continues to develop and continue to affect all parts of our civilization: government, educational institutions, medical, businesses and individuals.

Today educational institution and other commercial and non commercial organization cannot conduct its business without their dependence on complicated information technology infrastructures. Civilizations are competing to build information technology infrastructures to gain a competitive advantage. Organizations need these information technology infrastructures not only for their communication needs, but also for conducting there business activities.

In fact, there are quite a number of security issues in e-education system, for instance, user authentication and access control, non-repudiation for critical actions like course registration, examinations and assignment delivery, course tuition fee payment, confidentiality of user personal information, and course material copyright protection.

The security concerns may differ faintly depending on the type of courses offered by an organization. However, the most disturbing issue in e-learning system might be the copyright protection issue, which is essential to all kinds of electronic courses (e-course). This security issue may have the following picture:

One of the registered students violates the copyrights protection of the course materials by passing the course materials to other organization or to other non-registered students. Regularly, the organization that provides the course materials depends on the registration fee to keep up and maintain all activities and operations of the organization. Therefore, copyright protection violation rigorously exposes the income of the organization at risk.

New information and communication technologies have become major resources and basis for learning in higher education. Technologies have several potentials to support different instructional strategies and provide an efficient way of delivering e-course material and improving comprehension. The contemporary universities need to increase lifelong learning opportunities to its students any time, any place and at any rate to be successful in the global educational marketplace [1].

The use of e-leaning in the educational process has grown significantly in the last few years, however, it is a relatively insecure, hence, most educational organization haven't yet

taken into considerations or any new strategy for securing e-learning process [2]. Implementing e-learning is complex. Implementing e-learning is about project management, change management and risk and security management [3]. Additionally, the topic e-learning or e-education is having much attention especially because world-class universities such as MIT, Harvard and Stanford in the United States and Oxford in the United Kingdom are implementing it [4].

E-learning can be defined as the online delivery of information for purposes of education, training and knowledge management [5]. This definition means that the Internet and computer will be used in the e-learning process. Thus, e-learning is more complex and intertwined the opportunities for intrusion and attack. E-learning security involves more that just preventing and responding to cyber attacks and intrusion, it involves copyright protection, integrity, availability, non-repudiation, authentication and authorization.

The use of Internet application in higher education and in most organizations is being optimistic. The reasons are various and complex and lecturers in educational institution are under high pressure to learn and adopt this latest technology to support their teaching and their students' learning.

Using the Internet in the educational process is very beneficial to both students and educational institutions. The core reason why Internet are gaining so much interest lies on its ability of joining and interoperating heterogeneous communities. A lot of users who use different platform can communicate with each other easily on the Internet. The Internet and its potential and capabilities are very attractive; however, the current standards behind the technology need to be justified very carefully before deploying the Internet for very sensitive applications such as e-learning system. Since, default Internet transactions are unencrypted and unsecured, and they can establish the potential for disaster and failure [6].

Computer security is the shield that all types of organizations use to protect sensitive, commercial and classified information from unauthorized users. A break of this shield has implications that go far beyond any financial form that could be assigned to such an intrusion or adversary. The concepts of computer security are practically basic in nature, however, implementing security in a continually changing technological environment is a big challenge, but it has to be met by organizations, individual users and governments. Therefore, the threats in computer security must be understood.

This research presents the design and implementation of a global e-learning system that provides the basic security requirements including confidentiality, integrity, non-repudiation, replay protection and the most important entity authentication.

2. Cryptographic Techniques

Data communication is an important part of our living. Therefore, protection of data from misuse is essential. A cryptosystem defines a pair of data transformations called encryption and decryption. Encryption is applied to the plain text which is the data to be communicated to produce cipher text which is the encrypted data using encryption key. Decryption uses the decryption key to convert cipher text to plain text or the original data.

With strong encryption, computer users can send confidential contracts by email, or safely store corporate strategy on a notebook, or carry home spreadsheets on a floppy disk. The encryption software may even be free.

To improve the protection mechanism Public Key Cryptosystem was introduced in 1976 by Whitfield Diffe and Martin Hellman of Stanford University [7]. It uses a pair of related keys one for encryption and other for decryption. One key, which is called the private key, is kept secret and other one known as public key is disclosed to the public.

The message is encrypted with public key and can only be decrypted by using the private key. So, the encrypted message cannot be decrypted by anyone who knows the

public key and thus secure communication is possible. RSA [8] (named after its authors Rivest, Shamir and Adleman) is the most popular public key algorithm. In relies on the factorization problem of mathematics that indicates that given a very large number it is quite impossible in today's aspect to find two prime numbers whose product is the given number. As we increase the number the possibility for factoring the number decreases. Thus, we need very large numbers for a good Public Key Cryptosystem.

Authentication, confidentiality and data integrity can be addressed by studying cryptographic techniques [9]. In using such techniques, it is predictable that information in transmit through the Internet can bypass through various computers before it arrives its target. A malicious user of any of the intermediary computers can monitor the Internet traffic, eavesdrop, intercept, change or replace the data through its entire path. Cryptographic techniques can be used to protect these data. Encryption is the process that makes

information indecipherable (cipher text) unless having a decryption key [10]. It uses mathematical algorithms and processes to convert intelligible plain text to unintelligible cipher text and vice versa [11]. It can, therefore, reduce risks from an eavesdropping on a network.

Cryptography is one of the most important tools that enable networks and Internet applications because cryptography makes it possible to protect electronic information. The effectiveness of this protection depends on a variety of mostly unrelated issues such as cryptographic key size, protocol design, and password selection.

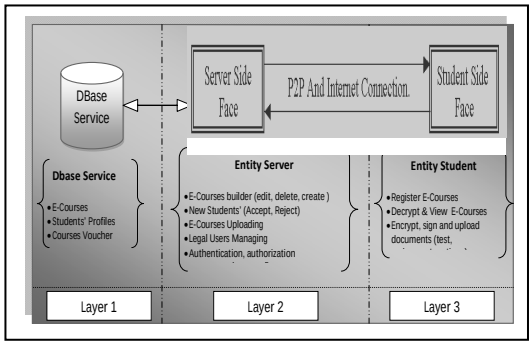
3. The Proposed Model (SeS)

The secure e-learning System (SeS) is a set of software modules designed so as to work together to protect the copyright of e-course material and to make the e-learning process more secure and trustee for organization and student a like. These modules are shattered amongst different components.

The proposed design model (SeS) follows the three-tier architecture model. This model breaks the software application into three different layers or tiers as shown in figure 1 below:

- Entity Student
- Entity Server
- Database services

Each layer has its own goals and design constraints and will be briefly explained in following sections.



1.1.1 Figure 1: Faces of SeS Model, own model

4. SeS Organization and Structure

The proposed model is organized to be employed in a traditional classroom using a LAN (Local Area Network) network or a WAN (Wide Area Network) connected using the standard TCP/IP protocol with an entity server representing the educational institution and an entity student representing the student workstation with a piece of software installed in it. The proposed system will start when an entity student communicate with the entity server to register as a new student using the educational institution Web site published on the Internet (see figure 2). The entity server, (see figure 3), will accept the student information and send to him, using his e-mail, an attached file contains a username and a password

that he/she can use to login into the site and download a software (Entity Student), (see figure 4), which he/she need to install into his computer so that he/she can use to register courses, add and remove courses and change his/her password. Additionally, with this software (Entity Student SW) the student can encrypt and decrypt files, view course materials and he/she can also send digital signed files using his/her public key to the instructor such as assignments, test, questions, etc.

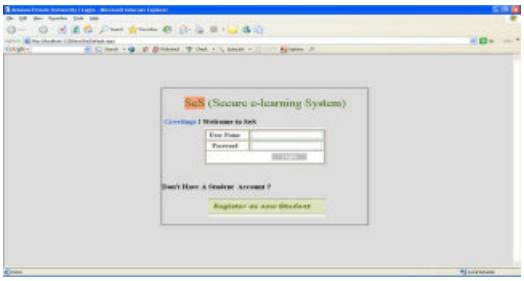


Figure 2: login and registration screen in the SeS system

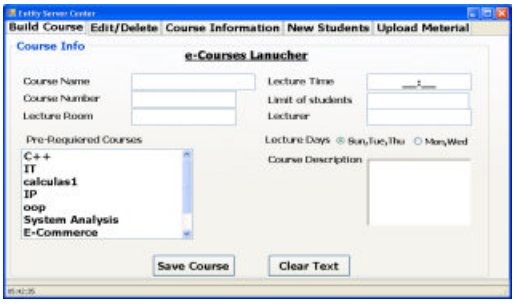


Figure 3: entity server solution

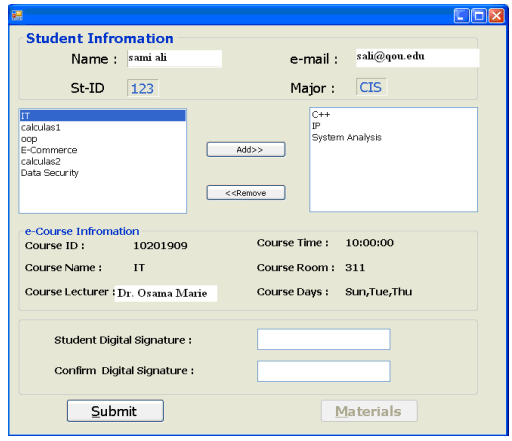


Figure 4: entity student software

Instructor in the educational institution could use a software called entity instructor to communicate with his/her students, send to them assignment, notes, receive from entity student assignment, check the digital signature of the entity student, and use this software to build and launch new e-course materials(see figure 5). However, when the entity student communicates with the entity server, an authentication scheme will be verified to insure security. These authentication scheme will include a precise time test, private and public key matching. When the entity server successfully completes the authentication scheme verifications, the entity student can be allowed to access the system and get what it requested. Otherwise the entity server will not allow the entity student to get through the system.

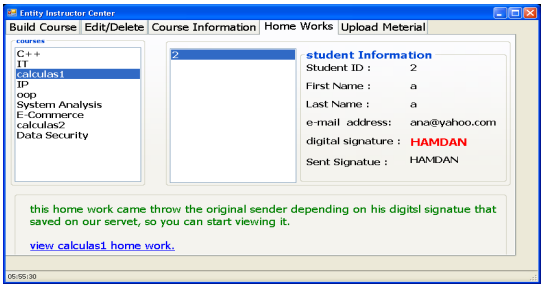


Figure 5: entity instructor software solution

This model, SeS, proposes a solution for the security problem of the e-learning system. The SeS based on the eXC model by Yau [2]. The eXC model proposed a solution for the copyright protection. The model uses the hardware configuration of the student's computer to protect the copyright of the organization's recourses materials. It supposes to provide a mechanism to protect the learning material from unauthorized distribution, and shows how this mechanism can be integrated in the operation model of online learning e-course providers. However, this model for Yau [2] is not fully protected. Hence, the student is able to do the copying or saving of the e-material using copy or save commands within the operating system. However, if the content of the e-Course includes many different files, the student might only be able to save one Web page at a time [12], and would have to call up the save function many times in order to get a complete copy of all the files of the e-Course. Or, better yet, the student or maybe an opponent may use some commands line based Web client that is able of downloading Web content recursively for example "wget" [13], which can significantly speed up the process of illegally copy the material.

This type of attack to a copyright protection system is very common, and can beat the system by creation illegal copies of the e-content. Therefore, this model that proposed by Yau need to be modified and optimized to prevent this type of making illegal copy of the e-content of the courses. Additionally, this model for Yau used the commercial SET (Secure Electronic Transaction) to perform all activities related to encryption and decryption, which make this model incomplete and ambiguous.

What the researcher is proposing will prevent student from using all commands used to save, copy, or move the contents of e-course materials. The student will not even select the text or used the right click button of the mouse, he/she wont be able to use all control functions such CTRL + C or CTRL + X. Additionally, with this proposed system every thing will be shown and briefly explain in minutes details, the encryption, decryption, protocol and all scheme used will be explained widely. The following sections contain more detailed description of the design and software implementation for each of the entities software that involved in the e-learning process.

5. The Entity Server (server side face)

The SeS model software consists of the server side application (entity server) that performs the business and education logic of the system. This software (see figure 2) is the core component of the SeS. It is responsible for performing the requirement for secure electronic transaction of e-course processing at the server level.

The modular approach for the design calls for the separation of work on dedicated servers each has its own functionality. Allocating the work in this approach assures the highest availability of resources and meets the scalability needs. Consequently, the educational institution environment consists of two essential servers: the institution web server and the institution database server. These two servers together form a logical entity which can be called the entity server. The Content of the entity server system is designed for the administrator and instructors to create and to launch e - course materials.

There is a SeS sub-module, called the Course Launcher, residing in the entity server which is used for launching e-course. The entity server is the administration center of the whole platform. It is used for handling student registration, course registration, course payment, managing encryption and decryption, authorization and authentications as well as course materials hosting and downloading.

The entity server is the piece of software that does the entire procedures and operations in the e-learning model, for example is accountable for entity authentication, care of all e-learning procedures and related rules. Therefore, the entity server needs are as follows:

1. entity server should be able to achieve entity authentication
2. entity server should have sufficient space memory to store the entire databases, queries and solutions.
3. entity server should give a time-stamp service to record the process
4. entity server should supply concurrent computer links by a wireline technology.

6. Course Voucher and Course Package Process

When the e-course launcher is used to launch the e-course materials, two objects will be created: the course package and the course voucher for each e-course created.

Each course had a course voucher, which contains related information to a specific course such as course name, course number and course contents (index), prerequisite etc. It also contains

an encryption key which can be used for decrypting the Course Package. This means that, for viewing the e-course material, student must have both the Course Package and the Course Voucher for the specific e-course. Once the e-Course is successfully launched, the Course Package will be distributed over the network and could be downloaded in an encryption form by entity student.

When legal entity student request to view a specific e-course, several processes have to be done:

1. The course launcher will send the course voucher encrypted using the private key (Kpr) of the specific course concatenation with the course package encrypted with a key to the entity server.
2. The course voucher will be encrypted using the private key of the course (Kpr) and will be stored in the courses database within the database service server.
3. The authorized entity student can download the course voucher, decrypt it using his public key and then can get the key for the specific e-course,
4. Entity student now can use this key to decrypt the course package and eventually view the course material offline using his own computer, (see figure 8).

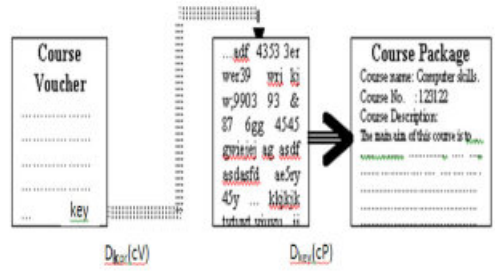


Figure 5: The Mechanism Of Viewing The E-course and Contents, own model

7. Digital Signature for Student

Each student completes the registration and the fees payment will be granted a public and private key (Digital Signature), which he/she can use to encrypt documents (Course Voucher, Courses Packages, announcements, courses details, grades, assignments etc.) to send them to the entity server or entity instructor and to decrypt documents, messages or files send to him by the entity server or entity instructor.

The RSA is a public-key cryptosystem has been used to present both encryption and digital signatures (authentication). Digital signatures are generated through entity server, as well as verified. Signatures are generated in conjunction with the use of a private key; verification takes place in reference to a corresponding public key. Each signatory (Registered student) has his own-paired public (assumed to be known to the general public) and

private (known only to the student) keys. Because an authorized student using his private key can only generate a signature, the corresponding public key can be used by anyone to verify the signature.

Therefore, a digital signature uses cryptographic technology to create an electronic identifier, but it can be used with any message, whether the message is encrypted or not. Thus, digital signatures can go together with an unencrypted or an encrypted message. Due to these criteria, a digital signature can be trusted and used like a written signature. For example, an entity student can use his digital signature with a private key that he keeps to himself. He then attaches this signature to a document and sends it to the entity server. His private key is mathematically linked to a public key that he posts on the entity server where his public key is stored. The recipient can then retrieve the sender's public key and reverse the process to determine the authenticity of the document.

The process for sending a digitally signed encrypted message is similar. In this case, the sender (entity server) must retrieve the entity student's public key from a public key database. Then uses it to encrypt the document and send it to the student. The recipient then uses his own private key to decrypt the document, and with this mechanism the entity server will be sure that only the recipient student can read it, thus, integrity, confidentiality and authentication will be assured. Additionally, the digital signature provides another advantage, the non-repudiation. In a cryptographic context, the word repudiation refers to the act of disclaiming responsibility for a message (ie, claiming it was sent by a third party). The mechanism strategy in the SeS model insist that the student attach a signature in order to prevent later repudiation, since the instructional organization may show the message to a third party to reinforce a claim as to its origin.

8. Student PC's License (STPCLicense.dll)

Potential student registers, fills the required information (Student Profile), pays fees and sends this information to the entity server using the Web site of the educational institution or any other secure communication channel. The entity server will receive and saved this information in the students' database.
Student now will be ready to register the course(s) needed according to his/her specification using the software installed in his PC. He could invoke the entity server to register the course. The entity server will immediately perform an authentication and authorization process. During the student registration process student's profile will be checked. A digital signature and a

private key will be added to the student's profile.

Through the installation, a public key-pair is generated using RSA scheme. A hardware profile copy the hardware (serial number of student PC's motherboard) configuration of the student's computer is also generated. The public key of the key-pair and this hardware profile are both stored inside a file called student PC's License (STPCLisence.dll). Besides, some personal information about the student is also stored in this student PC's License. This makes the STPCL unique to each computer. This License is then sent to the entity server. The entity server will verify this License, assign to it an expiry date, and sign it digitally. The server will send the signed License back to the student's computer. This copy of student PC's license will be stored during the student invocation of the server entity. This student PC's license will be checked when the student request the e-course material for viewing. All communication between the entity student and the entity server will be performed using encryption techniques to guarantee secure transferring of information between the two sides.

9. Requesting and Viewing e-Course

When an entity student invokes the entity server for viewing the course material, the student PC's License will first be examined and checked if this invocation is valid. The student will be allowed to access and have an encrypted copy of the e-course material if and only if the following conditions are satisfied:

1. The student PC's license file has not been expired yet.
2. The student PC's license file had properly signed by the server entity.
3. The software is invoked on the computer on which it was originally installed

During the invocation, a hardware configuration profile of the entity student's PC will be generated to test the current hardware configuration of the entity student's PC. This hardware profile is compared with the hardware profile that had been stored in the entity student PC's License. The third condition will only be meted if the two hardware profiles match.

When an entity student registered an e-Course, both the Course Package and the Course Voucher will be under the student's ownership. Since the encryption key to the Course Package is contained in the Course Voucher, the Course Voucher must also be protected on the student's computer. When the Course Voucher is received from the e-learning, it is encrypted with the computer's public key using the RSA asymmetric cryptographic algorithm. This public key is in fact the one stored in the students PC's License. After that, encrypted Course Voucher received from the entity server will be stored

encrypted in the entity student's computer. Since it is encrypted with the computer's public key, using the computer's private key can only decrypt it. The private key to this key-pair is stored in some special location on the student's computer hard disk.

If all conditions are met the entity student will be allowed to download the course material to their own computers, decrypt and view the material offline, while making it difficult to perform unauthorized copying (see figure 6 and 7).

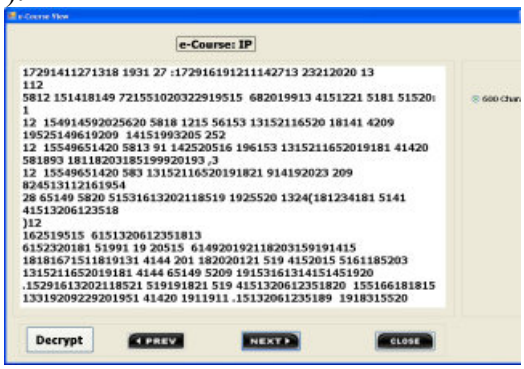


Figure 6: the encrypted e-course ready to be decrypted and viewed

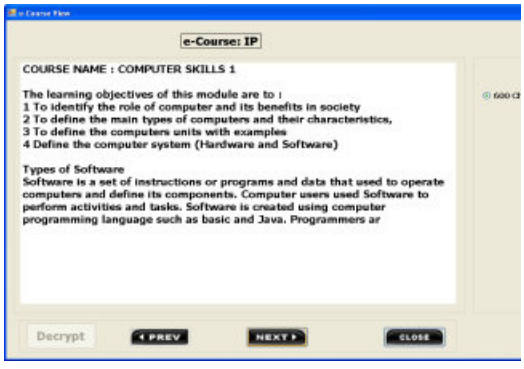


Figure 7: viewing the decrypted e-course material

Additionally, the Computer License is designed to have an expiry date, and the average lifetime of the students PC's License is six months. Before a Computer License expires, the entity server will keep track of students' Computer Licenses, it will make sure that there is only one valid student PC's License for each entity student. In a case where an entity student cheats and request for re-issuing the student License, or the student's private key is compromised, the old student License will be cancel, however, the student have to pay the registration fees once time again if he want to have another copy of the e-course material.

10. Online Submission of Assignments

The online learning system involves a variety of communication flows between the entity server and remote entity student, each of which may have different security requirements:

- general broadcasts (e.g. lectures, module material);
- student-specific (e.g. assignment, grades);
- submission (e.g. work for assessment);
- interactive (e.g. tutorials).

To make the communication between the entity student and entity server more

trustee for both side, each entity student will be granted a unique public key. This public key must be used by the entity student to digitally sign every document he/she sends to the entity server.

The SeS system will help students to solve and submit student homework assignments encrypted using their own private key. This will give great opportunity to e-education institute to force their students not to repudiate any document sent by them with their own private signature.

10.1 Digital Signature For Student

In the past, people perform their signature by signing a name or affixing a seal on a document to build the related rights and duties [14]. Hence, now we are lining in the age of Internet, e-commerce and e-government, the use of digital signature is very important.

- **Public Key Signature**

To provide for integrity, strong data authenticity and non-repudiation of all directory information is important to achieve some security features. In this way the student and organization can be sure that he/she is talking to the trusted directory when retrieving information. Digital signatures can be used to implement three important security services:

- Authentication – ensures that a principal is really who he/she claims to be.
- Data authentication – ensures that the data origin cannot be forged.
- Data integrity – ensures that no modification of data has been performed by unauthorized principals.
- Non-repudiation – ensures that a principal cannot deny performing some actions on the data (e.g. authoring, sending, and receiving).

Each entity student will be granted a digital signature that he must use to sign every document he sends to the entity server. This digital signature will be based on the RSA public key signature. Since entity student will be given two keys: public and private key. Using these two keys entity student will encrypt a secret word that is only known to him/her and chosen by him/her. This encrypted word will be sent to the entity server and be stored in the student database table to be checked and compared whenever entity student send each signed document. The following steps illustrate the public key signature scheme used in the SeS:

Suppose ES: entity server, EC: entity student, M: message

EC: (dEC, n)=Private key for Entity student

: (Xword) / entity student will select any secret word only known to him/

EC→ES: C= EKpr (Xword)

ES: M= DKpu(xword)

The above Public key signature can be explained as follows :

1. the RSA public key encryption will be used to generate a private key for the entity student (nEC , dEC) . Now student assume to recall a secret word that is only known to him.
2. Entity student send to entity server the secret word encrypted using his private key (dEC) so as to be used for verifying signature and documents or messages send by entity student.
3. Entity server will decrypt the secret word (XXX) using the entity server public key (eES) and store it the student database table.
4. Now whenever an entity student sends a signed document, his/her signature will be compared with his/her decrypted secret word which has been stored in the student database table.

Conclusion

E-educational institutions organizations have to present innovative approaches in its e-educational process. Effective adoption of e-education system has to be comprehensive and should include all aspects of security regards organizations and students alike

The e-learning security measures include the formulation and implementation of a policy for server security, configuration access control, users' access control and login passwords, in conjunction with public and private key cryptographic techniques, which used to achieve user authentication, provide a safeguard against attacks, and prevent non-repudiatory usage of system by legitimate students. These techniques in result will allow data integrity and confidentiality of the organization recourses

References:

[1] Schocken, S. (2001), "**Standardized frameworks for distributed learning**", *Journal of Asynchronous Learning Networks*, Vol. 5 No.2, pp.97-110.

[2] Yau, J.C.K., Hui, L.C.K., Cheung, B.S.N., Yiu, S.M., Cheung, V.L.S. (2002), "**A cryptographic schemes in secure e-Course eXchange (eCX) for e-Course workflow**", Conference Proceedings of SSGRR 2002 (Summer) International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet (SSRGG 2002s), L'Aquila, Italy, 29 July-4 August.

[3] Phillips, T., (1995) "**System Security in the National Information Infrastructure:**

Networks at Risk”, NCSA Conference Proceedings, April 1995.

[4] Efrain Turban, David King, Dennis Viehland, Jae Lee, 2006, “**E-Commerce A managerial Perspective**”, Prentice Hall, U.K, London.

[5] Allen, M. W. Michael Allen’s, 2003, “ **Guide to e-learning** “. Hoboken, NJ: John Wiley and Sons, 2003.

[6] David G. Rosado, Carlos Gutiérrez, Eduardo Fernández-Medina, Mario Piattini, 2006, “”**security patterns and requirements for internet-based applications**”, Volume 16 Number 5 2006 pp. 519-536 , Emerald Group Publishing Limited ISSN 1066-2243

[7] Needham, R.M., Schroeder, M.D. (1978), "**Using encryption for authentication in large networks of computers**", *CACM*, Vol. 21 No.12, .

[8] R.L. Rivest, A. Shamir, and L.M. Adleman, **A method for obtaining digital signatures and public-key cryptosystems**, Communications of the ACM (2) 21 (1978), 120-126.

[9] Needham, R.M., Schroeder, M.D. (1978), "**Using encryption for authentication in large networks of computers**", *CACM*, Vol. 21 No.12, .

[10] Chou, D. C. et al. (1999), “**Cyberspace security management”, Industrial Management & Data Systems**, Volume 99, Number 8, pp. 353-361, available at: **http://ejournals.ebsco.com/direct.asp?ArticleID=E35WHE69Q8AW23NFTVNX**

[11] RSA Security (2003), “**Understanding Public Key Infrastructure (PKI**)”, RSA Security Inc., available at: http://www.computel.com.lb/Downloads/PKI.pdf

[12] Joe Cho-Ki Yau1, Lucas Chi-Kwong Hui1, Siu-Ming Yiu1 and Bruce Siu-Nang Cheung, (2006), “**Towards a Secure Copyright Protection Infrastructure for e-Education Material: Principles Learned from Experience**”, international Journal of Network Security, Vol.2, No.1, PP.21–28, Jan. 2006

[13] GNU (wget), GNU wget,http://www.gnu.org/software/wget/wget.html.

[14] Sattar J. Aboud and Mohammad A. Al-Fayoumi, (2007), “ **A new Multisignature Scheme using re-encryption technique**” , Journal of Applied sciences, ISSN 1812-5654, Asian network for scientific information.

Automatic Essays Scoring (AES)

Hamzeh Mujahed, Labib Arafeh,
Al-Quds Open University , Al-Quds University

Abstract:

An Automated Essays Scoring (AES) system has been developed. The idea behind the proposed AES is to grade the essays by identifying the main keywords in the essays and their synonyms, and processing these keywords using modelling approach-based techniques including Fuzzy Logic, Clustering, and Neuro-Fuzzy. Currently, the developed AES can identify up to 15 keywords, each of which has up to 4 synonyms. A 100-word history essay has been used to test the AES. 1080-data sets have been constructed using 13 questions. The obtained average correlation coefficient between actual and predicted marks has a value of 0.9963 for training and 0.9937 for the testing data sets. Whereas, the Mean Absolute Percentage Error (MAPE) average value obtained is 0.0404 for the training and 0.0405 for the testing sets. These preliminary promising results demonstrate the adequacy of adopting the modelling techniques in solving the automated scoring systems. Further investigation is currently accomplished to take the order of words and negations issues into account.

Key words: *AES, Fuzzy Inference, Scoring, Neuro-Fuzzy*

1 Introduction

Automated Essay Scoring (AES) can be defined as a computer technology that evaluates and scores the written prose (Dikli ,2006). AES systems are now appearing in the educational institutions, and are increasingly being accepted as a way of efficiently grading large numbers of essays (Williams, 2006).In educational institutions, when large numbers of students' answers are submitted at once, teachers find themselves bogged down in their attempt to provide consistent evaluations and high quality feedback to students within as short a timeframe as is reasonable. The efficiency AES holds a strong appeal to institutions of higher education that are considering using standardized writing tests graded by AES for placement purposes or exit

assessment purposes(Wang ,et al, 2007)

The growth of e-Learning systems has increased greatly in recent years due to the demand by students for more flexible learning options and economic pressures on educational institution, which see technology as a cost saving measure.One of the major aspects of developing e-Learning systems is how to assess students knowledge based on essay type answers(Oriqat, 2007).The result of growth in e-Learning systemss led to number of studies conducted to assess the accuracy and reliability of the AES systems with respect to writing evaluation. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006).

2 Related work

A number of studies have been conducted to assess the accuracy (measurement of the degree of agreement between actual marks and predicted marks) of the AES systems with respect to writing assessment. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006).

One of the main studies at A-l-Quds University (Oriqat, 2007 concentrated on using fuzzy logic to score the short essays based on short answers by determining five main keywords and synonyms (inputs). These inputs have been processed by developing models based on fuzzy and Neuro-Fuzzy approaches. The obtained result from the models was promising and showed high agreement between actual and predicted marks. Our Fuzzy Automated Essays Scoring System (FAESS) is represented in fig. (4.1). we have pre-process stage to determine the fifteen main keywords and synonyms necessary for the

systems to predict the mark for longer answers. The main difference between the two approaches relies on number of keywords which are important factors to deal with longer answers. Also in our work we have concentrated in scoring the essays on content dimension and we have explored the importance / impact of words' order in the sentence and we have also explored negation's issue in the sentence, and ways to solve.

3. Fuzzy Inference System (FIS)

Fuzzy Inference Systems are currently being used in a wide field of applications. In recent years, fuzzy modeling technique have become an active research area due to its successful application to complex system model, where classical methods such as mathematical and model-free methods are difficult to apply because of lack of sufficient knowledge (Priyono, 2005). One popular approach is to combine fuzzy systems with learning techniques derived from neural networks; such approaches are usually called Neuro-fuzzy systems (Singh, et al, 2005). For the most complex system where few numerical data exist and only ambiguous or imprecise information may be available, fuzzy reasoning provides a way to understand system behavior by allowing us to interpolate approximately between observed input and output situation. Reasoning based on fuzzy approaches has been successfully applied for inference of multiple attributes containing imprecise data.

There are two most used types of Fuzzy Inference System (FIS): Mamdanis' and Sugenos'. These two types of inference systems vary somewhat in the way the outputs are determined. The general formula for the rules in our developed Mamdani type are:

IF (KW_i is MF_j) and (KW_{i+1} is MF_j) and ….. and (KW_m is MF_j) THEN (Mark is MF_k) …….. *(1)*

Where i = 1 to m represent the i^{th} keyword. m = 15, number of keywords

MF_j is the J^{th} membership function where j=1 to 7; and k= 1 to 17 represent the k^{th} output membership

function for the predicted mark, and KW is the abbreviation for keyword or one of its synonyms.

For Sugeno FIS, it is similar to the Mamdani method in many respects, the main difference between Mamdani and Sugeno is that the Sugeno output is usually a linear function . A typical rule in a Sugeno fuzzy model has the form:

$$R_i : IF\ (KW_1\ is\ A_{i1})\ and \ldots and\ (KW_m\ is\ A_{im})\ THEN\ Y_i = a_{i1}KW_{1+} \ldots + a_{im}KW_{m+} a_{i0} \quad \ldots\ldots\ldots (2)$$

Where R_i $(I = 1, 2, \ldots, c)$ denotes the i^{th} fuzzy rule, are the input (antecedent) variables,

Y_i are the rule output variables, A_{i1}, …, A_{im} are fuzzy sets defined in the antecedent space, and

a_{i1}, …, a_{im}, a_{i0} are the model consequent parameters that have to be identified in a given data set. Fig. 3.1 below show the general block diagram for developed models.

4. The Developed Models

The general block diagram in Fig. 4.1 shows the general architecture for the AES developed models. In the pre-process stage, we have

defined the questions and its reference answers, identified the system constraints, and determined the main keywords and synonyms to be ready for the input to the fuzzy system.

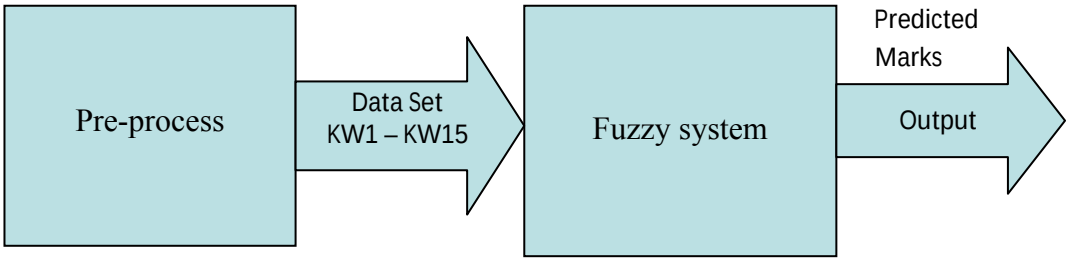


Figure (4.1) AES general block diagram

In the following section, three models based on fuzzy and Neuro-fuzzy have been constructed on 1080 data set collected from students answers related to historical topic.

4.1 Multiple Input Single Output (MISO) Mamdani Model

The MISO model have fifteen input , each input represent one main keyword or its synonyms and each input have number of membership functions were each function correspond to a weighting value from an answer document that are suitable to the input.

Table 4.1 shows the results obtained for training and testing data for each answer set. The average results for all data set also calculated. Some results obtained for testing data set are better than the results obtained from trained data that is because we have training a general model for all sets.

Table 4.1 :MISO Mamdani model results

Question No.	Training/Testing answers	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9901	0.128	0.069	0.994	0.0688	0.075
2	46/24	0.99	0.088	0.809	0.9923	0.106	0.069
3	40/20	0.99	0.1393	0.0796	0.995	0.1182	0.11
4	53/27	0.996	0.0921	0.063	0.9928	0.184	0.087
5	73/37	0.934	0.115	0.178	0.9934	0.264	0.095
6	120/60	0.995	0.074	0.04	0.993	0.2406	0.0715
7	67/33	0.985	0.022	0.0528	0.9378	0.0410	0.1998
8	33/17	0.948	0.0319	0.116	0.9795	0.0229	0.1045
9	33/17	0.9959	0.1267	0.085	0.9877	0.049	0.1159
10	40/20	0.993	0.207	0.083	0.9928	0.1291	0.1477
11	53/27	0.994	0.097	0.0656	0.9777	0.1932	0.1906
12	33/17	0.994	0.0969	0.084	0.9901	0.1029	.1159
13	60/30	0.987	0.021	0.05	0.9539	0.0345	0.1782
Average		**0.984**	**0.0953**	**0.13653**	**0.9830**	**0.119554**	**0.120008**

The results obtained in table 4.1 are promising and the average correlation between predicted and actual mark approximately more than 0.95 which best describe the agreement between actual and predicted marks. Fig. 4.2 shows a sample of the agreement plot between actual and predicted marks related to one of the questions (TR6 data set).

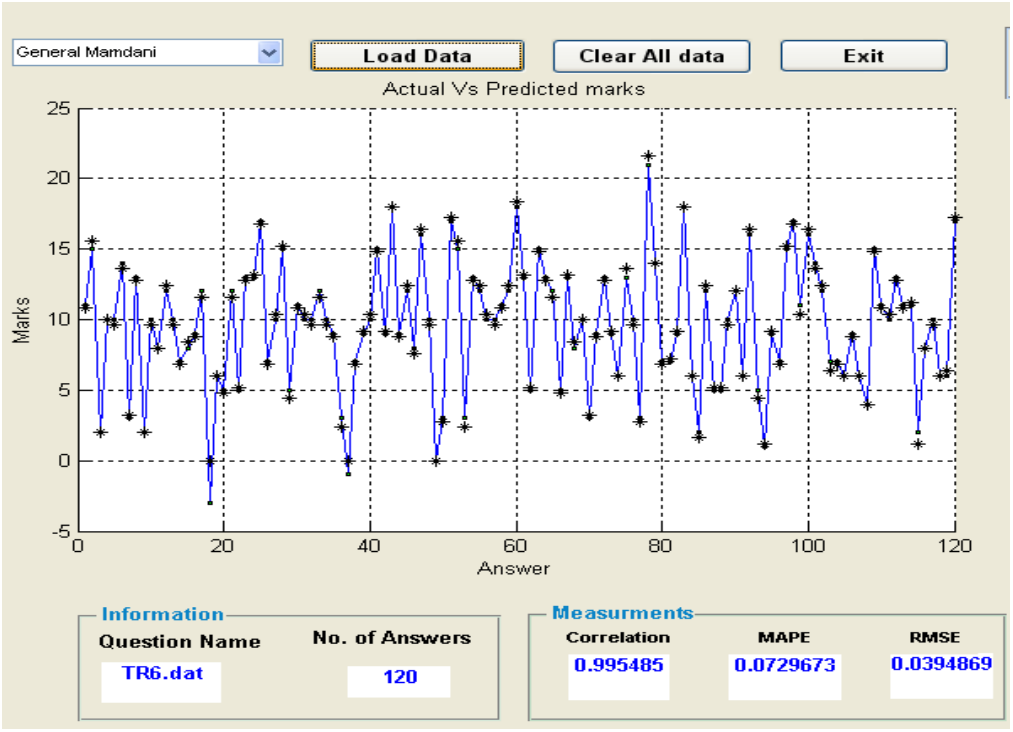


Figure (4.2): MISO Mamdani Model plots for TR6 dataset

The stars in fig.4.2 represent the actual marks; whereas the squares with line connected each square represent the predicted marks. Fig. 4.3 shows the correlation measurement between training and testing data.

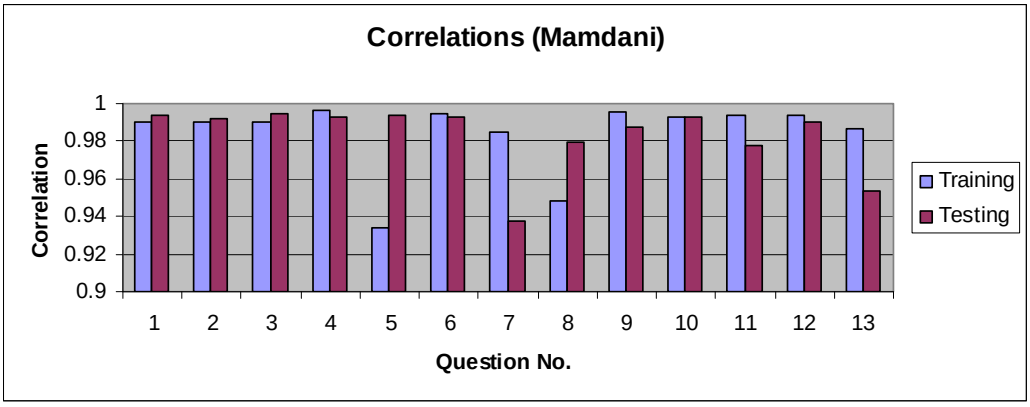


Figure (4.3): Correlation results for MISO Mamdani model

4.2 Grid Partition Sugeno with back propagation optimization Model

The difference between mamdani and sugeno FIS lie in the consequent of the fuzzy rules and hence the agrregation and defuzzification procesure accordingly. Table 4.2 shows the results obtained for training and testing data for each answer set. The average results for all data set also calculated

Table 4.2: Grid partition Sugeno model results

Question No.	Training/Testing	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9952	0.0829	0.048	0.9966	0.0381	0.0572
2	46/24	0.9939	0.0798	0.0654	0.9945	0.0987	0.0811
3	40/20	0.9664	0.341	0.2074	0.9973	0.0628	0.084
4	53/27	0.9969	0.0743	0.0569	0.9972	0.0879	0.0543
5	73/37	0.976	0.1943	0.1093	0.9835	0.2738	0.1504
6	120/60	0.9954	0.0729	0.0394	0.997	0.1088	0.0467
7	67/33	0.9914	0.0181	0.0399	0.9945	0.0197	0.0599
8	33/17	0.9879	0.0193	0.0569	0.9883	0.0181	0.079
9	33/17	0.9931	0.1443	0.1102	0.9927	0.0388	0.0892
10	40/20	0.9961	0.1269	0.0632	0.9975	0.0551	0.0861
11	53/27	0.9931	0.1047	0.074	0.9819	0.2033	0.1718
12	33/17	0.9959	0.0941	0.0731	0.9218	0.3295	0.3213
13	60/30	0.991	0.0179	0.042	0.9933	0.0209	0.0682
Average		**0.9901**	**0.1054**	**0.0758**	**0.9873**	**0.1042**	**0.1037**

Fig. 4.4 shows the correlation measurement between training and testing data. The preliminary result looks promising with high correlation values of an average value 0.9873. This in turns indicates the high performance for the developed model.

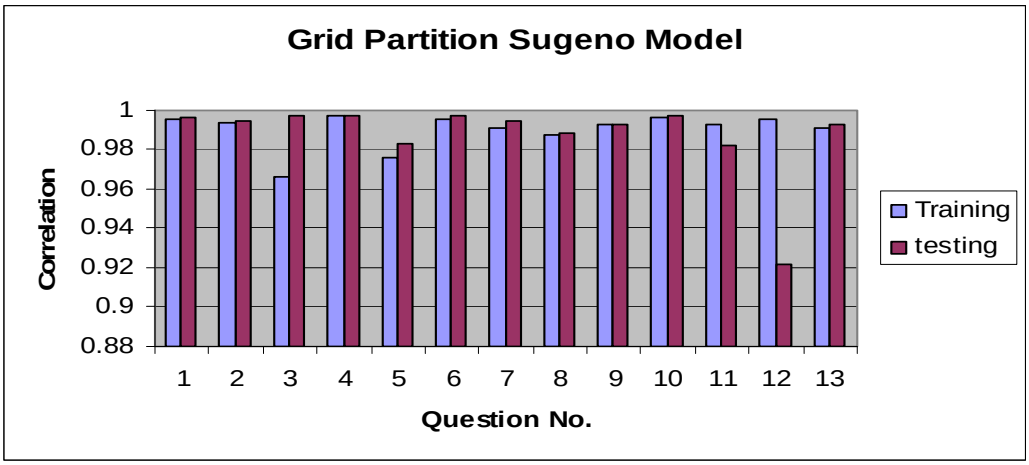


Figure (4.4): Correlations for Grid partition Sugeno model

4.3 Sugeno Sub-clustering model

The purpose of subtractive clustering is to identify natural grouping of data from a large dataset to produce concise representation of a systems behavior. The clustering model was build using 1080 dataset for training and testing the model. Table 4.3 shows the results obtained for training and testing data for each answer set

Table 4.3: Sugeno sub-clustering model results

Question No.	Training/Testing	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9968	0.0659	0.039	0.9976	0.0292	0.048
2	46/24	0.9949	0.0728	0.0593	0.9973	0.0349	0.0563
3	40/20	0.9984	0.039	0.0447	0.9985	0.0339	0.0622
4	53/27	0.9985	0.0349	0.0392	0.9978	0.0663	0.048
5	73/37	0.9976	0.036	0.0344	0.9985	0.0547	0.044
6	120/60	0.998	0.0324	0.0258	0.9979	0.079	0.0386
7	67/33	0.9936	0.0165	0.344	0.9598	0.0299	0.1614
8	33/17	0.9894	0.0185	0.0534	0.9898	0.0172	0.0739
9	33/17	0.9985	0.047	0.0507	0.9941	0.0347	0.08
10	40/20	0.998	0.0620	0.0449	0.9984	0.0465	0.0682
11	53/27	0.9978	0.0428	0.0414	0.9977	0.0487	0.0615
12	33/17	0.998	0.0417	0.0511	0.9964	0.0336	0.0695
13	60/30	0.9934	0.0163	0.0361	0.9954	0.0187	0.0568
Average		**0.9963**	**0.0404**	**0.0664**	**0.9937**	**0.0405**	**0.0668**

Fig. 4.5 shows the correlation measurement between training and testing data. An agreement value (Correlation) shows a value of 0.9963 for trained data subset and 0.9937 for untrained (testing) data.

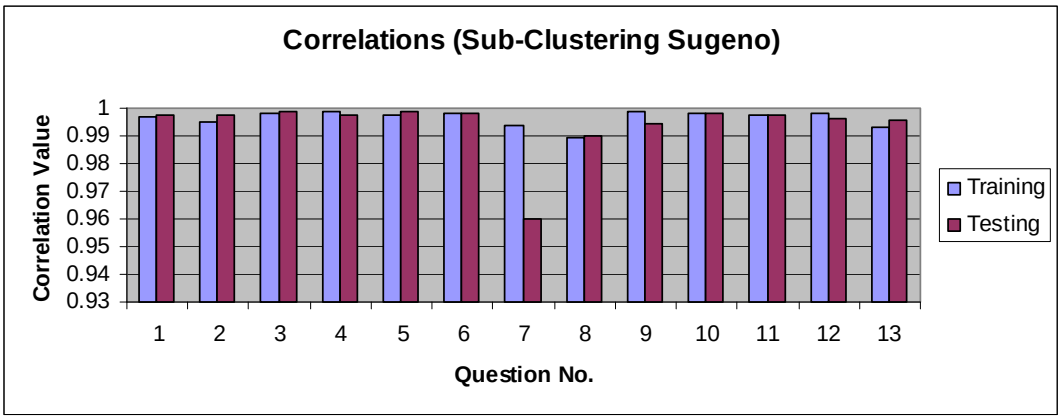


Figure (4.5): Correlations results for Sub-clustering Sugeno model

To investigate further in the development of models and to improve the results obtained, we cascade more than model to produce hybrid model. When we cascade two models, the average correlation obtained for training data is approximately equal the correlation for other models, that's because the average correlation for our developed models are high . The results of cascade two models are very good and show high agreement (correlation) between actual and predicted marks.

5 Discussions

We have developed three basic models based on fuzzy and Neuro-fuzzy system to train and test the AES system. The preliminary results obtained are promising in general. The correlation value for the thirteen answers dataset (Question1 to question 13) of the 1080 sets are clearly used to check the models. The graph represented in Fig. 5.1 shows that Sugeno Sub-clustering technique produced the best results, while MISO Mamdani model produced a very good results but have the least performance compared to the other developed models.

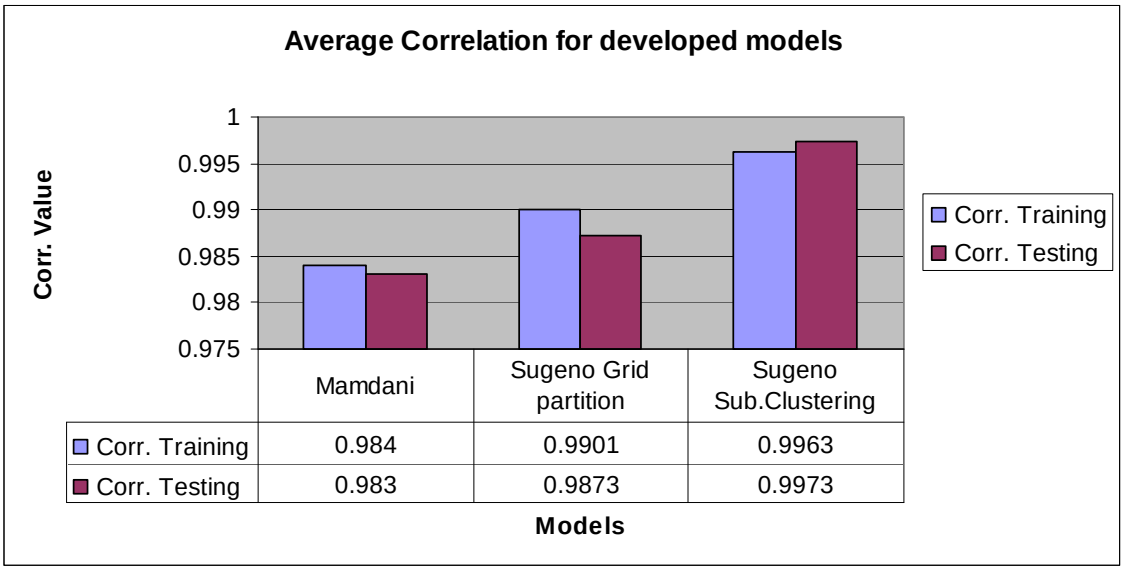


Figure (5.1): Average Correlation for the developed models

6 Conclusions

The work on this paper concentrated on developing a system for AES purpose. Therefore, we have explored the integrated and adaptive Neuro-fuzzy approach. The developed AES based on input, process and output. The input is the assessed subject that is related to historical subject, the output will be the predicted marks. we used FIS and neural learning approaches to develop our three models. The comparison between our three models using average results of correlation, RMSE, and MAPE shows that using Sugeno sub-clustering model produced the best result. The preliminary results obtained from our models are promising and shows the capability to adopt it in AES systems.Further testing and comparisons with other similar AES systems will be accomplished and published.

Currently, we are enhancing our developed models to take the negation and the order of words into accounts. Further more, an online AES system will be uploaded and tested by several colleagues each with his/her own essay.

References:

[1] Burstein, J., Chodorow, M., Leacock, C., Criterion^{SM} Online essay Evaluation: An Application for Automated Evaluation of Student Essays.

[2] Dikli ,S., An Overview of Automated Scoring of Essays, The Journal of Technology, Learning, and Assessment, Volume 5, Number 1 ,August 2006.

[3] Hearst, M., The Debate on Automates Essay Grading, University of California, Berkeley, California, USA,October,2000.

[4] Jyh-Shing ,Jang, R., ANFIS: Adaptive-Network-Based Fuzzy Inference system, Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, May 1993**.**

[5] Mellor ,A., Essay Length- Lexical Diversity and Automatic Essay Scoring ,Department of Media Science,Faculty of Information Science and Technology, September 30, 2010.

[6] Oriqat, Y. ,Modeling Techniques Applied to short Essay Auto-Grading problem, Al-Quds University, Jerusalem, Palestine, 2007.

[7] Palmer , J., William, R. , Dreher, H. (2002) , Automated Essay Grading System Applied to a First Year University Subject – How Can We do it Better? , Curtin University of Technology, Perth, WA, Australia.

[8] Persing, I., Davis, A. and Vincent N., Human Language Technology Research Institute, University of Texas at Dallas, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 229–239,MIT, Massachusetts, USA, 9-11 October 2010.

[9] Priyono,A., Generation of Fuzzy Rules with subtractive clustering, Universiti Teknologi Malaysia, 2005.

[10] Singh, T. N., Kanchan, R. Verma , A. K Saigal, K. A comparative study of ANN and Neuro-fuzzy for the prediction of dynamic constant of rockmass,India, February 2005

[11] Wang ,J. , Brown, M., S., Automated Essay Scoring Versus Human Scoring: A Comparative Study, The Journal of Technology, Learning, and Assessment, Volume 6, Number 2 · October 2007

[12] William, R., The Power Normalized Word Vectors for Automatically Grading Essays, School of Information Systems, Curtin University of technology, Perth, Australia, Volume 3, 2006.

[13] Yen-Yu C. et al , Intelligent Systems magazine , Volume: 25 Issue: 5 , Sept. 2010.

Author(s):

[1] Hamzeh, Mujahed, Academic Supervisor

Al-Quds Open University, Department Information Technology and Communication

Hebron, West Bank, Palestine

Email: hmujahed@qou.edu

[2] Labib, Arafeh, Associate Professor

Al-Quds University, Department of Graduate Studies, Faculty of Engineering

Abu dis, Jerusalem, West Bank, Palestine

Email: larafeh@eng.alquds.edu

Electric Power Load Short Term Forecasting

Rae'd Basbous, Dr. Labib Arafeh
Al-Quds Open University, Palestine
Al-Quds University, Palestine
rbasbous@qou.edu , larafeh@eng.alquds.edu

Abstract

At this study, two kinds of models have been developed, namely the Single Input Single Output, SISO, and Multiple Input Single Output, MISO. These two developed approaches depend on the Fuzzy based techniques including integrated and adaptive Neuro-Fuzzy approaches, and have been compared to represent the Short Term Load Forecasting, STLF, models.

Different models for SISO and MISO have been developed using the training data, such as, Sugeno Fuzzy Inference System with different optimization techniques including Hybrid and Back-propagation optimization techniques, Sugeno model using the Subtractive Clustering, and finally Sugeno cascaded model using Subtractive Clustering and Hybrid optimization technique.

The developed models have been integrated with a stand-alone application with Graphical User Interface, GUI. The developed Electric Power Load Forecasting System, EPLFS, can be accessed online to predict the power load.

The preliminary and promising results indicate the suitability and adequacy of the developed models depending on the Fuzzy approach to solve the short term load forecasting using the time and weather variables

Keywords—ANFIS, STLF, Subtractive Clustering, Sugeno.

1. Introduction

Load forecasts (LF) is an important component of power system operation and planning involving prognosis of the future level of demand to serve as the basis for supply-side and demand side [1]. Precise load forecasting helps the electric utility to make unit commitment decisions, reduce spinning reserve capacity and schedule device maintenance plan properly [2].

LF can be divided into three main categories according to [3]. These categories are Long-Term Load Forecasting, LTLF, Mid-Term Load Forecasting, MTLF, and Short-Term Load Forecasting, STLF.

STLF cover an interval ranging from an hour to a week [3]. For STLF several factors should be considered, such as time factors, and weather data. STLF is important for different functions such as unit commitment, economic dispatch, energy transfer scheduling and real time control.

In this research, different modeling techniques for the short term load forecasting problem have been explored. Different measures have been used to check the adequacy of the developed models, these measures include the

Correlation Coefficient (CC), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).

Two kinds of models have been developed, namely the Single Input Single Output, SISO, and Multiple Input Single Output, MISO. These two developed approaches depend on the Fuzzy based techniques including integrated and adaptive Neuro-Fuzzy approaches, and have been compared to represent the STLF models. Real historical data profiles for two years (2006 and 2007) have been used to develop and test these proposed models. These data profiles were provided by Jerusalem District Electric Company (JDECO), and the Palestinian Meteorology Office (PMO).

A pre-processing stage has been accomplished for the collected data. The bad data and outliers in the collected data identified and removed. In order to develop and test the models, we have divided the data using a cross validation algorithm to training and testing datasets (75% of the available historical data profiles have been used for training and 25% for testing).

In Developing the SISO models the time parameter was considered as an input, while in MISO models three inputs have been considered namely, Time, High and low Temperatures. For the two kinds of models, the power load was the output.

Different models for SISO and MISO have been developed using the training data, such as, Sugeno FIS with different optimization techniques including Hybrid and Back-propagation optimization techniques, Sugeno model using the Subtractive Clustering, and finally Sugeno cascaded model using Subtractive Clustering and Hybrid optimization technique to improve the models outcome. These models have developed using the Fuzzy toolbox in Matlab 2008.

These models have been integrated with a stand alone application with GUI called "Electric Power Load Forecasting System, EPLFS". This application has been developed using the Matlab 2008 GUI toolbox. Load forecasting can be done using this system, so one can load datasets saved in a text file, obtain the forecasted load using the developed models, plot the predicted power loads and the actual ones if known, and evaluate the predicted output by calculating the various measures used including the CC, MAPE and RMSE.

The EPLFS has been tested using the obtained power load historical profile for the year 2008 and used as the actual load. The system has been used to predict the load for one day and one week ahead using the developed models.

The organization of his paper is as follows. Section II describes the data provided. In addition, the analysis and of data and preprocessing is also presented. Section III presents and reviews the currently available fuzzy-based modeling techniques. Section IV covers the implementation and development steps that were followed to explore the use of soft computing approaches for

STLF. The results of the developed models presented in section V. Section VI summarizes and concludes the paper.

2. Data Description

2.1 Data Sources

Developing any supervised-based soft computing model needs pairs of data (inputs and output), and in order to have a reliable STLF models that best represents the trends of these input and output data, we need reasonable actual sets of data composed of the electric power load as an output for a certain time during a day with known weather conditions as an inputs for a specific line that provides a chosen area.

The sources of the available datasets profiles are Jerusalem District Electric Company (JDECO) and Palestinian Meteorology Office (PMO). The provided power historical data profile includes the time and the corresponding power load at that time, while the weather historical data profile includes humidity, highest temperature, and lowest temperature for each day.

2.2 Data Preprocessing

The selection of the training datasets from the available data significantly affects the forecasting results, and to achieve a reliable and a more comprehensive approach to load forecasting, the days which have similar load and historical temperature values should be chosen to train (develop) the models [4].

Many factors affect the success of model training on a given task. The representation and quality of the instance data is first and foremost [5]. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult.

In our research, pre-processing stages have been accomplished for the collected data. These stages are as shown in Fig. 1. They are, obtaining historical profiles, input variables selection, bad data and outliers detecting and removing, time formatting, and cross validation.

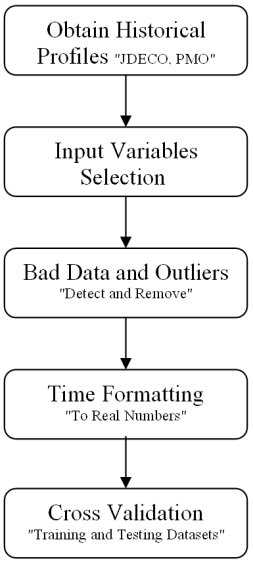


Fig. 1: Data Preprocessing Procedure Stages.

In the second stage, the time, power load, and temperature elements (low and high temperatures) have been selected to train the models. After that an existing algorithm for detecting and removing the outliers from the historical profiles has been used, and a manual procedure has been followed for detecting and removing the bad data such as the zero loads. Time formatting in the fourth stage is necessary since the input to the models should be a real number format and not in a time format (hour: minutes) as provided to us. Finally, a cross validation technique has been applied to divide the datasets into training and testing datasets.

Fig. 2 depicts a sample for detecting the outliers for (two months 7/2006, 7/2007) which was used to develop our July models before removing the outliers.

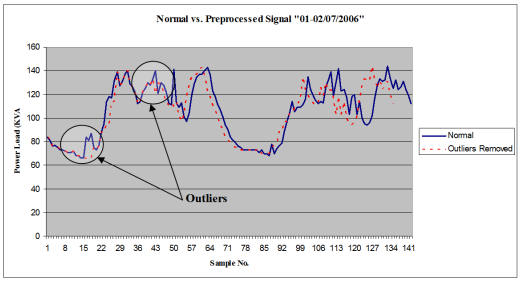


Fig. 2: The Original Dataset Before and After Removing the Outliers.

The figure shows the original signal for the first two days of the datasets (01-02/07/2006). Three clear outliers which have been detected and removed were circled.

2.3 Extra Testing Datasets from the Year 2008

In order to test our developed models using new unseen datasets, new profiles have been obtained for the year 2008 from JDECO and PMO for using the developed models to predict the power load for one day and one week. A small two samples selected to test the developed models. The first one is for one week from July (01-07/07/2008) to test the general July and Summer MISO models. The second datasets are for one week from May (0-07/05/2008) to test the general May and Spring MISO models.

3. Neuro Fuzzy Inference Systems

3.1 Fuzzy Inference Systems

Fuzzy Inference Systems (FISs) are also known as fuzzy rule-based systems [3], fuzzy model, fuzzy expert system, and fuzzy associative memory. This is a major unit of a fuzzy logic system. The decision-making is an important part in the entire system. The FIS formulates suitable rules and based upon the rules the decision is made.

The basic FIS can take either fuzzy inputs or crisp inputs, but the outputs it produces are almost always fuzzy sets. When the FIS is used as a controller, it is necessary to have a crisp output. Therefore in this case defuzzification method is adopted to best extract a crisp value that best represents a fuzzy set.

FIS consists of a fuzzification interface, a rule base, a database, a decision-making unit, and finally a defuzzification interface. A FIS with five functional block described in Fig. 3.

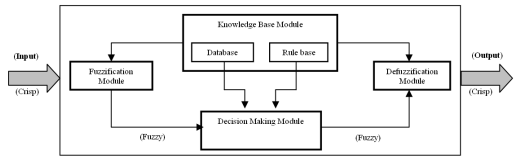


Fig. 3: Fussy Inference System [10].

The most important two types of fuzzy inference method are Mamdani's fuzzy inference method, which is the most commonly seen inference method.

This method was introduced by Mamdani and Assilian in 1975 [6]. Another well-known inference method is the so-called Sugeno or Takagi–Sugeno–Kang method of fuzzy inference process. This method was introduced by Sugeno et al. in 1985 [7]. This method is also called as TS method.

A typical fuzzy rule in a Sugeno fuzzy model has the format [8]:

$$IF\ x\ is\ A\ and\ y\ is\ B\ THEN\ z = f(x, y) \quad (1)$$

Where AB are fuzzy sets in the antecedent; $Z = f(x, y)$ is a crisp function in the consequent. Usually $f(x, y)$ is a polynomial in the input variables x and y, but it can be any other functions that can appropriately describe the output of the system within the fuzzy region specified by the antecedent of the rule.

For STLF a typical rule in a MISO Sugeno fuzzy model with three inputs (Time, HiTemp, LowTemp) and one output (PowerLoad), has the form [8]:

$$If\ Time\ is\ Time_j\ and\ Hi\text{-}Temp\ is\ HiTemp_k\ and\ Low\text{-}Temp\ is\ LowTemp_l,\ then$$

$$PowerLoad = p_i\ Time_j + q_i\ HiTemp_k + r_i LowTemp_l + s_i \quad (2)$$

Where ($_j$) represent the time input MF, ($_k$) represent the high temperature input MF, and ($_l$) represent the low temperature input MF. The terms p_i, q_i, r_i, s_i, indicate the consequent parameters which determined through the training process.

3.2 Adaptive Neuro Fuzzy Inference Systems

In fuzzy modeling, the membership functions and rule base are generally determined by trial-and-error approaches. Although this approach is straightforward, the determination of best fitting boundaries of membership functions and number of rules are very difficult. In order to calibrate the membership functions and rule base in fuzzy modeling, the neural networks have been employed by researchers [9]-[14]. This system has been called fuzzy neural, neuro-fuzzy or adaptive network based system. The key properties of neuro-fuzzy systems are the accurate learning and adaptive capabilities of the neural networks, together with the generalization and fast-learning capabilities of fuzzy logic systems.

ANFIS constructs a fuzzy inference system (FIS) whose membership function parameters are adjusted using either a backpropagation algorithm alone, or in combination with a least squares method. This allows the fuzzy systems to learn from the data they are modeling. The purpose of ANFIS is to integrate the best features of Fuzzy Systems and Neural networks.

The Least-Squares (LSQ) optimization algorithm [15]-[17] is a mathematical optimization technique that attempts to find a function which closely approximates a given dataset. It tries to minimize the sum of the squares of the ordinate differences between points generated by the function and corresponding points in the dataset.

The Hybrid Learning (HL) algorithm [10] and [18], which combines the Gradient Descent (GD) and the LSQ algorithms, is one of the widely used algorithm in the literature to identify the parameters of the ANFIS. In the HL algorithm procedure, there are two passes which are called forward pass and backward pass. In the forward pass, functional signals go forward until the de-fuzzy layer and the consequent parameters are identified by the LSQ algorithm. In the backward pass, the error rates propagate backward and the premise parameters are updated by the GD algorithm.

Back-propagation learning algorithm, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. It is a very powerful method to adjust the weights of the neural network It was first described by Paul Werbos in 1971 which he published in his doctoral thesis [19], but it wasn't until 1986, through the work of [20], when Rumelhart et al, rediscovered this technique, that it gained recognition, and it led to a "renaissance" in the field of artificial neural network research.

3.3 Data Clustering

Data Clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [18]. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis.

Four of the most representative off-line clustering techniques frequently used in [3]:

1. K-means (or Hard C-means) Clustering,
2. Fuzzy C-means Clustering,
3. Mountain Clustering, and
4. Subtractive Clustering.

Mountain Clustering, proposed by Yager and Filev [18]. This technique calculates a mountain function (density function) at every possible position in the data space, and chooses the position with the greatest density value as the center of the first cluster. It then destructs the effect of the first cluster mountain function and finds the second cluster center. This process is repeated until the

desired number of clusters has been found.

Subtractive clustering, proposed by Chiu [18]. This technique is similar to mountain clustering, except that instead of calculating the density function at every possible position in the data space, it uses the positions of the data points to calculate the density function, thus reducing the number of calculations significantly.

4. Implementation and EPLFS Development

4.1 Implementation

In system modeling and identification, the important steps are to identify structure and parameters of the system based on the available data. The structure identification itself can be considered as two types, identification of the input variables of the model and the input–output relation. Most of the modeling approaches consider the input variables as a known priori and hence only the input and output relation has to be found [23].

In order to deeply study the effect of the temperature in the Short Term Load Forecasting (STLF), two kinds of models have been developed, SISO and MISO models. For the SISO models that shown in Fig. 4 the time has been used as the input for the models and the power load at that time has been used as the output.

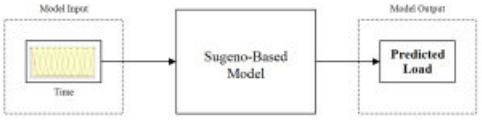


Fig. 4: SISO Sugeno FIS Model Architecture

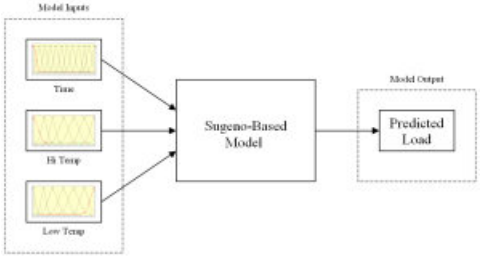


Fig. 5: MISO Sugeno FIS Model Architecture

In the MISO models which shown in Fig. 5, three variables (Time (T), High Temperature (HT), and the Low Temperature (LT) for that day) have been used as an input for the developed models, and the power load at that time was considered as the model output.

The results in Table I show that the input variable temperatures have a considerable effect on load forecasting. The results in the table are for the two kinds of models, SISO and MISO Sugeno models with hybrid optimization. The adequacy of the two developed models has been checked using the CC and two error measures; MAPE, and the RMSE.

Table I: Temperature Effect on Load Forecasting.

Type of Model	No. of MFs	Dataset	Errors Measures		
			CC	*MAPE*	*RMSE*
MISO	12 7 7	Training	0.9815	0.0257	0.0769
		Testing	0.9719	0.0324	0.1715
SISO	12	Training	0.9106	0.0667	0.1662
		Testing	0.9085	0.0656	0.3045

The proposed models are to be trained with the obtained historical data profiles from JDECO and PMO before testing them. The first step for training a model is obtaining an accurate historical data. In addition, data should be chosen that is relevant to the model. Several models have been developed such as, Sugeno with different optimization techniques, Subtractive Clustering, and finally Subtractive Clustering cascaded with Hybrid optimization technique to improve the developed models.

Fig. 6 illustrates a general developing "training" block diagram of our models. It consists of three main stages.

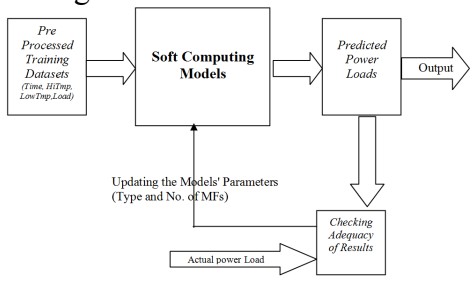


Fig. 6: A General Block Diagram for Developing/Training Soft Computing Models

These stages can be summarized as follow:

1. The first stage is pre-processing the input datasets for the system. These datasets include four elements; three of them are inputs; namely, the time, the high temperature, and the low temperature of the

day and one output (the actual loads).

2. The second stage is concerned with various soft computing models that have been developed. SISO/MISO models will be developed using different techniques (Hybrid and Back-propagation optimization techniques, Subtractive Clustering and finally by cascading two models). The same datasets (Training and Testing) have been used in developing all the models.
3. The third stage checks the adequacy of the developed models to demonstrate their performance.

Three measures according to Basbous in [3] have been used to effectively check the adequacy of results, and these measures illustrated bellow:

1. *The CC measure between actual and predicted power loads*. It indicates the strength and direction of a linear relationship between the forecasted and actual loads and calculated by [8]:

$$CC_{xy=}\sqrt{1-\frac{\sum_{i=1}^{N}(y_i-x_i))^2}{\sum_{i=1}^{N}(y_i-y)^2}} \quad ,(3)$$

where y_i : is the ith actual data,
y : is the average of all actual data,
x_i : is the ith predicted data.
N : is the number of data points under consideration.

2. *The Mean Absolute Percentage Error (MAPE)*, which has been traditionally used to measure accuracy in load forecasting [2]. It captures the proportionality between the forecast error and the actual load. The MAPE is calculated by [2]:

$$\text{MAPE}= \sum_{i=1}^{N}\left|\frac{y_i-x_i}{y_i}\right|*\frac{100}{N}\% \quad ,(4)$$

3. The Root Mean Square Error (RMSE), which is used to evaluate the error (differences) between the forecasted and actual loads. The general form of the RMSE equation for the actual power loads (Y) and the predicted ones (X) in given by [24]:

$$\text{RMSE}= \frac{\sqrt{\sum_{i=1}^{N}(y_i - x_i)^2}}{(N-1)} \quad ,(5)$$

In order to obtain the best results from the developed models, the model parameters (type and number of membership functions) need to be updated and determined manually to be fixed for the all models as shown in Fig. 6. The predicted loads will be measured against the actual marks and the system parameters will be altered to find the best outcome. We have been tried several Membership Functions (MFs) including: Gaussian Curve, Generalized Bell, Trapezoidal and Triangular. Table II lists the results that obtained form a July SISO model with different MFs and the same parameters (No. of MFs, hybrid optimization technique, training datasets and testing datasets).

Table II: Results for a July SISO Model with Different Types of MFs.

Type of MF	No. of MFs	Dataset	Training Dataset Errors		
			CC	*MAPE*	*RMSE*
Gaussian Curve	12	Training	0.9104	0.0667	0.1664
		Testing	0.9094	0.0654	0.3032
Trapezoidal		Training	0.9100	0.0669	0.1667
		Testing	0.9088	0.0656	0.3041
Triangular		Training	0.9094	0.0672	0.1672
		Testing	0.9079	0.0661	0.3055
Generalized Bell		Training	0.9106	0.0667	0.1662
		Testing	0.9085	0.0656	0.3045

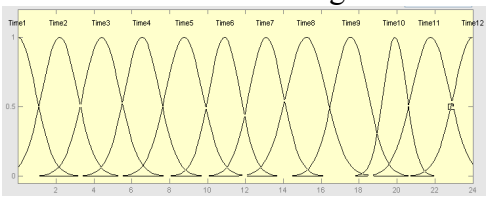
As listed in Table II, there is no major difference between the outputs (predicted loads) regarding the type of the membership function. However, the MF that produced the best results is found to be the Generalized Bell (GBell).

Table III lists the results that obtained form a July MISO model with different number of MFs and the same parameters (Type of MFs (GBell), hybrid optimization technique, training datasets and testing datasets).

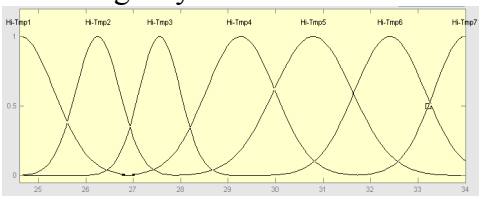
Table III: Results for a July MISO Model with Different Number of MFs.

No. of MFs	Dataset	Errors		
		CC	*MAPE*	*RMSE*
6 6 6	Training	0.9252	0.0593	0.1526
	Testing	0.9307	0.0570	0.2666
8 6 6	Training	0.9511	0.0466	0.1242
	Testing	0.9528	0.0449	0.2212
12 6 6	Training	0.9521	0.0452	0.1229
	Testing	0.9511	0.0448	0.2250
12 7 7	Training	***0.9815***	***0.0257***	***0.0769***
	Testing	***0.9719***	***0.0324***	***0.1715***
12 8 8	Training	0.9561	0.0428	0.1179
	Testing	0.9548	0.0440	0.2167

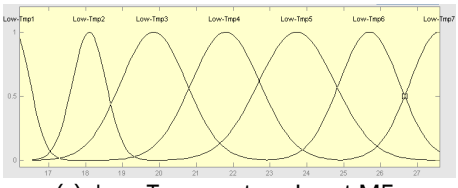
It is clear from Table III that the best results obtained when we have been used 12 MFs for the time input, and 7 MFs for the temperature inputs (High and Low). According to the results shown in Table III, we will fix the number of the MF's in the proposed models to 12 MF's for the time input and 7 MF's for the Low and High temperatures inputs. Fig. 7 represents the inputs MFs that we have been used in building a MISO model using July datasets.



(a): Time Input MFs



(b): High Temperature Input MFs



(c): Low Temperature Input MFs

Fig. 7: MISO Model Inputs (Time, High Temperature, Low Temperature) MFs

Table IV shows the results obtained from a two July MISO Sugeno model with hybrid optimization the first one developed using all the training datasets available.

Table IV: Results for a July MISO Model Before and After Removing the Outliers.

Datasets Used	Dataset	Errors		
		CC	*MAPE*	*RMSE*
All the Datasets	Training	0.9414	0.0510	0.1336
	Testing	0.9255	0.0568	0.2585
Outliers Removed	Training	0.9815	0.0257	0.0769
	Testing	0.9719	0.0324	0.1715

The second one developed using the training datasets after removing the outliers. The results obtained show the effect of the outliers to the accuracy of the models. The model that has been developed before removing the outliers shows a CC value between the actual and predicted loads for the training datasets of 0.9414 while the CC value for the model that has been developed after removing the outliers is 0.9815.

Using the parameters that give us the best results and the outliers have been removed from the datasets, two SISO and MISO models have been developed using datasets from the month of July and May. In addition to that, two SISO and MISO models have been developed using datasets from Spring (month of April and May) and Summer (month of July and August) seasons. Several models with different optimization techniques have been developed for these types of models.

4.2 Stand Alone Application " EPLFS"

These models were integrated within a stand alone application with GUI. The developed Electric Power Load Forecasting System "EPLFS" is as shown in Fig. 8; the figure demonstrates the forecasted power loads for a testing datasets. The three lists in the system present the times, forecasted loads, and the actual loads. Three different measures appear in the bottom of the right corner, the CC, MAPE, and RMSE.

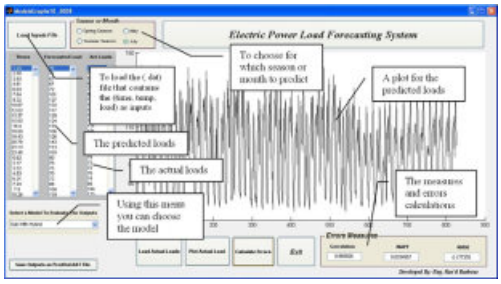


Fig. 8: The EPLF System: Showing a Plot for the Predicted Loads in a Certain Hours.

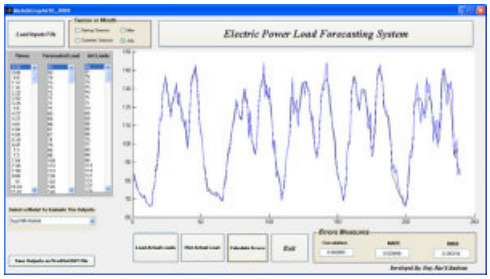


Fig. 9: The EPLF System: Showing a Plot for the Predicted Loads vs. Actual Ones.

Using EPLFS we can load the datasets, evaluate the predicted output using the developed models, plot the actual and predicted load, and calculate several measures including the CC, MAPE and RMSE.

Fig. 8 represents a snapshot for the EPLFS when used to predict the power loads for certain times with temperatures parameters.

Fig. 9 represents a snapshot for the EPLFS when used to predict the power loads for certain times with temperatures parameters. A plot for the actual and predicted power loads are seen in the snapshot. The blue dotted line represents the actual loads and the black continuous one represents the predicted power loads.

To load an input and actual load files to the EPLFS, it should be in (.DAT) format. For the input file it should be with three columns. The first column representing the times formatted as mentioned in the previous chapter. The second and third columns representing the temperature parameters (High and Low) respectively. For the file that containing the actual loads to be compared with the predicted ones, the loads should be in one column.

5. Results and Comparisons

In this research and using the available historical datasets profiles, we have been started by developing eight Sugeno models with hybrid optimization technique. Four of the models are SISO models and the other four are MISO

models. These models have been used to predict the power load in specific months (May, July) or general model to be used in seasons (Spring, Summer). Another eight models have been developed using the same datasets but with the back-propagation optimization technique. After that the Subtractive clustering has been used to construct a Sugeno models for the same months and seasons. Then, the Subtractive clustering with the Hybrid optimization technique have been used to construct a cascaded model to improve and enhance the results obtained from the previous models.

As mentioned before, the training data sets and the models parameters (number and type of MF, number of rules, and cluster radius) have been fixed and used for the proposed models. For example 12 MF for the Time input and 7 MF for the Low and High temperatures of the type G-Bell have been fixed for all the models with Hybrid and Back-propagation optimization techniques. Furthermore, for the models that have been constructed using the Subtractive Clustering, a cluster radius of the value 0.1 has been fixed and used to construct these models.

5.1 Results of the Developed Models

Table V below lists the average CC measures for all the developed models using Hybrid, Back-propagation, Subtractive Clustering and Subtractive clustering with Hybrid optimization (cascaded model). It is clear from the table that the cascaded model has produced the best results with an average equal to (0.97) for the MISO models followed by the models that have been developed using the subtractive clustering with an average correlation equal to (0.95). The results of the cascaded models have been obtained using less number of rules compared with the models that have been developed using the Hybrid and Back-propagation optimization techniques.

Table V: The Correlations Measures Average for all the Developed Models

Optimization Method	SISO		MISO	
	Training	*Testing*	*Training*	*Testing*
Hybrid	0.87770	0.87323	0.95483	0.94928
B-Prop	0.82280	0.81780	0.90940	0.90448
Subtractive	0.87880	0.87428	0.95925	0.94813
Sub. with Hybrid	0.87935	0.87345	0.97238	0.96158

The developed cascaded model with cluster radius equal to 0.1 produced 313 rules, while with Hybrid and Back-propagation optimization techniques the number of rules that have been produced equal to the multiplication of inputs membership functions (12 MFs for the time * 7 MFs for the Hi-Temp * 7 MFs for the Low-Temp) which is equal to 588 rules.

In addition to the CC, two different error measures, the MAPE and the RMSE have been used to examine and show the adequacy of the developed models and its outcome. The CC measures the agreements between the actual and predicted power loads, while the error measures RMSE and MAPE give an indication how the performance of the developed models are.

Table VI shows the average results of the error measure MAPE for all the models. It is clear from the table that the lowest values are for the MISO models developed using the cascaded models with an average of the MAPE equal to (0.03). This low value reflects the highest CC that achieved from these models as shown in the Table V. You can notice that the MISO models have the lowest MAPE values over the SISO models because of the temperature parameters effect on the power load. The same thing has been noticed in the CC measures since the MISO models produced the best results and have the highest CC values.

Table VI: The Average MAPE Measures for all the Developed Models

Optimization Method	SISO		MISO	
	Training	*Testing*	*Training*	*Testing*
Hybrid	0.08685	0.08793	0.05068	0.05410
B-Prop	0.10453	0.10510	0.07778	0.07735
Subtractive	0.08618	0.08738	0.04623	0.05335
Sub. with Hybrid	0.08603	0.08763	0.03878	0.04668

Table VII shows the average results of the error measure RMSE for all the models. This measure show the adequacy of the developed models too in addition to the MAPE measures. The same thing for the RMSE results as in the MAPE results achieved where the cascaded model have the best results (the lowest RMSE values). The developed MISO models with the temperature

parameters produce the lowest RMSE values over the SISO models. It is clear from the table and the figure that the lowest values are for the MISO models developed using the cascaded models with an average of the RMSE equal to (0.08).

Table VII: The Average RMSE Measures for all the Developed Models

Optimization Method	SISO		MISO	
	Training	*Testing*	*Training*	*Testing*
Hybrid	0.17580	0.31280	0.10505	0.19883
B-Prop	0.20608	0.36710	0.15453	0.27873
Subtractive	0.17463	0.31113	0.09908	0.18628
Sub. with Hybrid	0.17423	0.31240	0.08310	0.16485

A graphical representation for the average CC and error measures (RMSE and MAPE) are shown in Fig. 10. The three measures are combined together in the same graph to take a clear look to the behavior of these measures. A relation can be concluded from the graph which is: an increasing in the CC leads to decrease in the error measures (RMSE and MAPE). As an example for this, from the above tables when the average CC for the MISO cascaded models are equal to (0.97) the corresponding average MAPE and RMSE values for the cascaded model are equal to (0.03 and 0.08) respectively. The second case is when the average CC measure for the SISO cascaded models is equal to (0.87) the corresponding average error measures MAPE and RMSE for the cascaded model are equal to (0.08 and 0.17) respectively.

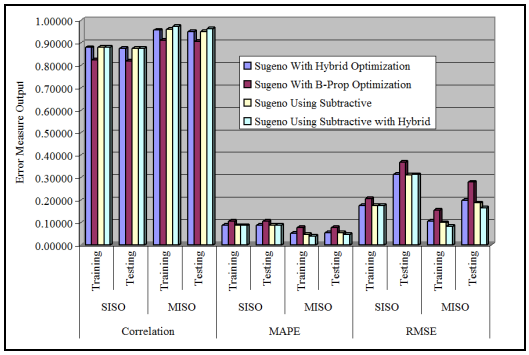


Fig. 10: The Average Correlation Measures Against the Error Measures (MAPE and RMSE) for all the developed SISO and MISO models

Fig. 10 can be summarized by the following points:

1. The developed cascaded models (Subtractive Clustering with Hybrid optimization) produced the highest results. The average CC between the actual and predicted loads ranging between 0.95 and 0.98 with average equal to 0.97.
2. The developed models with Back-propagation optimization techniques have the lowest CC compared to the other models and similarly for the two error measures. The correlation coefficient ranging between 0.89 and 0.91 with average equal to 0.90.
3. The developed models using the Subtractive Clustering can be enhanced when subject to the Hybrid optimization technique. The average correlation for the developed models using the Subtractive Clustering is equal to 0.95 and after applying the Hybrid optimization technique it is enhanced to an average equal to 0.97.
4. Finally, we can notice that the highest the CC the lowest the MAPE and RMSE. For example, the average CC of the cascaded model is equal to 0.97, the average MAPE is equal to 0.03, and the average RMSE is equal to 0.08. While, the average CC of the developed models using the Back-propagation optimization technique is equal to 0.90, the average MAPE is equal to 0.07, and the average RMSE is equal to 0.15.

5.2 One Day and One Week Ahead Prediction Using Unseen Datasets from the Year 2008

As mentioned in chapter four, new historical profiles have been obtained from JDECO and PMO for the year 2008. The selected datasets (one day and one week from July and May) have been used to test the EPLFS for new unseen

datasets. These datasets have not been considered in the cross validation process that has been applied in developing our models. The first day of May and July (01/05, 01/07) has been chosen in order to use the system to predict the load for one day. To test the system for predicting the load for a period of one week the first week of May and July (01-07/05, 01-07/07) has been considered.

These datasets have been not used in the cross validation for developing the models. The average CC that has been obtained for one day prediction is equal to 0.94 and the average MAPE is equal to 0.058. The average correlation in case of one week prediction is equal to 0.93 while the average MAPE is equal to 0.059. Table VIII and Table IX show the average correlation and error measures for one day and one week power load prediction using the developed models.

From these two tables it is clear that the best results have been obtained from the general May model in case of one day and one week prediction. The lowest average CC has been obtained from the general Spring model because of the wide range of the model input parameters (the wide variations of low and high temperatures). These results show the accuracy of the developed models to predict the power loads for the new unseen datasets one day and one week ahead.

Table VIII: The Average CC and Error Measures for One Day Prediction

Model	*CC*	*MAPE*	*RMSE*
July	0.9680	0.0407	0.2242
Summer	0.9435	0.0506	0.3187
May	0.9794	0.0350	0.1902
Spring	0.8738	0.1078	0.4874
Average	***0.9412***	***0.0585***	***0.3051***

Table IX: The Average CC and Error Measures for One Week Prediction

Model	*CC*	*MAPE*	*RMSE*
July	0.9428	0.0533	0.2872
Summer	0.9255	0.0576	0.3513
May	0.9578	0.0464	0.2850
Spring	0.9047	0.0809	0.4308
Average	***0.9327***	***0.0595***	***0.3385***

Fig. 11 and Fig. 12 below show the results obtained from the EPLFS for one day and one week prediction using the developed models.

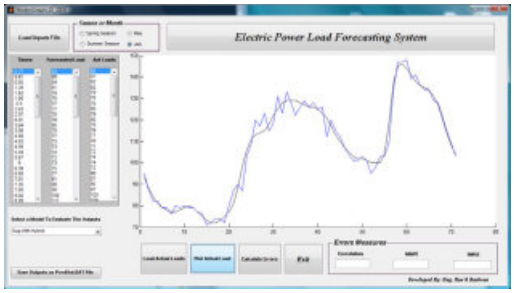


Fig. 11: One Day Prediction using the EPLFS for New Unseen Datasets.

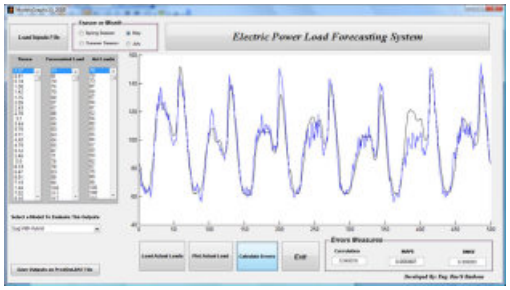


Fig. 12: One Week Prediction using the EPLFS for New Unseen Datasets.

5.3 Comparison with Other Studies

A plenty of works can be found in the STLF field. Some of these works are mentioned here briefly. Many papers that have been published recently in the refereed journals are considered and the ones whose main interests are STLF by soft computing methods are taken into account.

It is important to mention that different datasets and different approaches have been used in developing these models. However, we are trying to compare the obtained average CC and MAPE as a measure of errors. Furthermore, the same equations of the CC and MAPE that have been presented in (3) and (4) used in these papers to check the adequacy of the developed models. All the CC and MAPE results listed as a real number with fractions instead of using the percentage sign.

Hwang [25] described a new practical knowledge-based expert system (called LoFY) for short-term load forecasting equipped with graphical user interfaces. Also, various forecasting models like trending, multiple regression, artificial neural networks, fuzzy rule-based model, and relative coefficient model have

been included to increase the forecasting accuracy. The simulation based on historical sample data shows that the forecasting accuracy is improved when compared to the results from the conventional methods. Through the fuzzy rule-based approach, the forecasting accuracy at special days has been improved remarkably. The average MAPE results found for this system is 0.020.

Khan [1] presented a comparative study of six soft computing models namely multilayer perceptron networks, Elman recurrent neural network, radial basis function network, Hopfield model, fuzzy inference system and hybrid fuzzy neural network for the hourly electricity demand forecast of Czech Republic. The soft computing models were trained and tested using the actual hourly load data obtained from the Czech Electric Power Utility for seven years (January 1994 – December 2000). A comparison of the proposed techniques is presented for predicting 48 hourly demands for electricity. Simulation results indicate that hybrid fuzzy neural network and radial basis function networks are the best candidates for the analysis and forecasting of electricity demand for the experimented data, with the following MAPEs: For weekday forecast, 0.010 by radial basis function networks, 0.009 by fuzzy neural network; and for weekend forecast, 0.013 by radial basis function networks, 0.020 by fuzzy neural network.

A modeling technique based on the fuzzy curve notion is proposed by Papadakis [26] to generate fuzzy models for STLF. Different forecast models are developed for each day type in every season. The model is considered as a fuzzy neural network described in terms of a parameter vector and is trained using a genetic algorithm with enhanced learning and accuracy attributes. The performances of the developed fuzzy models are tested using load data of the Greek interconnected power system. They achieve a MAPE of 0.0167 with the data from year 1995.

A feed-forward neural network with a back-propagation algorithm is presented by Bhattacharyya [27] for three types of short-term electric load forecasting: daily peak (valley) load, hourly load and the total load. The forecast has been made for the northern areas of Vietnam using a large set of data on peak load, valley load, hourly load and temperature. The data were used to train and calibrate the artificial neural network, and the calibrated network was used for load forecasting. The results obtained from the model show that the application of neural network to short-term electric load forecasting problem is very useful with quite accurate results. The method has given the best performance with 0.9427 average CC and 0.108 average MAPE.

A Fuzzy Logic (FL) expert system is integrated with Artificial Neural Networks (ANN) for a more accurate short-term load forecast is presented by Tamimi [28]. The 24 hour ahead forecasted load is obtained through two steps.

First, a FL module maps the highly nonlinear relationship between the weather parameters and their impact on the daily electric load peak. Second, 12 ANN modules are trained using historical hourly load and weather data combined with the FL output data, to perform the final forecast. Comparisons made between this model, an ANN model, and an Autoregressive Moving Average (ARMA) model were show the efficiency and accuracy of this new approach. The average MAPE for these methods is found equal to 0.029.

From the results mentioned above, it is clearly noticed that the soft computing methods provide a promising solution to the STLF problem. In addition, combining or integrating more than one method together leads to an enhancement to the proposed models. For example Tamimi [28] combined the NN with FL, and [25] developed a forecasting system and include it with different forecasting models to increase the forecasting accuracy. Furthermore, the lowest results over the proposed models that haven mentioned above found in the models with Back-propagation optimization proposed by Bhattacharyya [27]. This agrees with the results obtained from our developed Sugeno FIS models with Back-propagation optimization.

Comparing our results with these systems we can see that our developed models produced satisfactory results using the temperature parameters only to predict the electric load without taking in account the other weather parameters or the type of the day or any other conditions. In our developed models, the CC for one day ahead prediction for the unseen datasets from the year 2008 ranges between (0.87 and 0.97) with an average value 0.94; the corresponding MAPE ranges between (0.03 and 0.10) with an average 0.05. Whereas; the obtained CC for one week ahead prediction for the same datasets ranges between (0.90 and 0.95) with an average value 0.93; the corresponding MAPE ranges between (0.04 and 0.08) with an average 0.05.

An improvement to the results that have been obtained from our developed models can be achieved when a hourly temperature and weather data profiles are available. Also, an average high and low temperature of the day which has been considered in the MISO models leads to reduce the error measures between the actual and predicted power loads compared to SISO models.

6. Summery and Conclusion

The general objective of this work is to explore the use of soft computing and artificial intelligence approaches to develop Short Term Load Forecasting (STLF) system that predict the power load for one day up to one week ahead in specific month or specific season. The introduction of NF modeling approaches to the area of load forecasting has been presented. The basic concepts of

ANFIS, Hybrid Learning, BP Learning, and Data Clustering have been reviewed in the previous sections. The implementation of and NF approaches to model the relationships between Temperature, Time, and Power Load has been introduced.

Real JDECO power line in Bier Nabala village and PMO historical data profiles for two years (2006 and 2007) have been collected and used to develop and test the various models. Developing the models for load forecasting has been applied firstly using the available datasets. Two kinds of models have been developed, Single Input Single Output (SISO) models, and Multiple Inputs Single Output (MISO) models. Three main inputs (Time (T), High Temperature (HT), and Low Temperature (LT)) and one output (Power Load (PL)) have been considered in building these models. For the SISO models, only the time has been considered as the input for the models and the power load at that time has been used as the output.

It has been found that the temperature is a major input parameter on STLF. Models, that do not utilize temperature measurements in training, produce quite larger errors than the ones exploiting them as input parameters. The correlation has been improved from 91% for the SISO model to about 98% for the MISO model as demonstrated in Table (5.1) using the same parameters (number of MFs, type of MFs, and cluster radius). These results clearly reveal the effect of the temperature parameters on predicting the power load.

Fuzzy Inference System (FIS) with different optimization techniques have been used to develop our models. Firstly we started by developing a SISO and MISO models using ANFIS with hybrid optimization technique. Then a SISO and MISO models have been developed using Back-propagation optimization technique. After that the Subtractive clustering has been used to develop such models. Finally cascaded models have been developed for STLF by constructing the models using the Subtractive clustering and then learning these models using the hybrid optimization techniques to achieve more accurate models.

While testing these models using the testing datasets that has been isolated before the training stage using the developed cross validation algorithm, the average obtained CC between the actual and predicted power loads for all the developed SISO models ranges between (0.81 and 0.87), with an average CC value 0.85; The corresponding average MAPE that ranges between (0.08 and 0.10) with an average value of 0.09, and average RMSE that ranges between (0.31 and 0.36), with an average value of 0.32. Whereas; the obtained average CC for the MISO models ranges between (0.90 and 0.96) with an average CC value 0.94; The corresponding average MAPE that ranges between (0.04 and 0.07) with an average value of 0.05, and average RMSE that ranges between

(0.16 and 0.27) with an average value 0.20. This demonstrates the adequacy of adopting these types of approaches to STLF problem as well as the improvements of models' forecasting performance when taking the weather into consideration.

It was noticed that the forecasting performance has been furtherlly improved by the MISO cascaded models, while maintaining all other factors including MFs types and numbers, and cluster radius. This improvement is notices as an improvement in the obtained CC that results from the MISO cascaded models which ranges between (0.95 and 0.97) and with an average CC value of 0.96 for all the developed models as compared with the forecasting performance of CC that ranges between (0.90 and 0.94) when developing the models using the other optimization techniques; The corresponding MAPE ranges between (0.03 and 0.06) with an average value 0.04, and RMSE ranges between (0.07 and 0.20) with an average value 0.16 compared to average MAPE ranges between (0.05 and 0.07), and average RMSE ranges between (0.18 and 0.27) for the other optimization techniques.

These models have been integrated with a stand alone application with GUI. The developed Electric Power Load Forecasting System "EPLFS" can be accessed online through either a local area network, or using a web server. The EPLFS has been tested using the obtained power load historical profile for the year 2008 and used as the actual load. The system has been used to predict the load for one day and one week ahead using the developed models. The CC for one day ahead prediction ranges between (0.87 and 0.97) with an average value 0.94; The corresponding MAPE ranges between (0.03 and 0.10) with an average 0.05, and RMSE ranges between (0.19 and 0.48) with an average 0.30. Whereas; the obtained CC for one week ahead prediction ranges between (0.90 and 0.95) with an average value 0.93; The corresponding MAPE ranges between (0.04 and 0.08) with an average 0.05, and RMSE ranges between (0.28 and 0.43) with an average 0.33.

Finally, different works in the field of STLF using different techniques accomplished by other researchers have been compared with our developed models. These works show the ability of the soft computing techniques to represent the STLF, and agree with our results that the Back-propagation optimization technique produced the lowest results. Our over all results indicates the suitability and adequacy of the developed models to solve the short term load forecasting problem using the time and weather variables.

ACKNOWLEDGMENT

The authors wish to thank Eng. Mansour Nassar and Eng. Salah Alqam from the strategic planning department in Jerusalem District Electricity Company (JDECO) for opening up the load consumption database for this scientific research and also to Mr. Isam Al-Seifi from Palestinian Meteorology Organization (PMO) for providing the temperature measurements.

References

[1] Khan, M., Abraham, A., Ondrusek, C. (2001), 'Soft Computing for Developing Short Term Load Forecasting Models in Czech Republic', Hybrid Information Systems, pp. 207-222, 2001.

[2] McSharry, P. (2006), 'Evaluation of Short-Term Forecasting Methods for Electricity Demand in France', RTE-VT workshop, Paris, May 29-30, 2006.

[3] Basbous, R. (2009), 'Electric Power Load Short Term Forecasting', M.Sc. Thesis, Al Quds University, 2009.

[4] Peng, M., Hubele, N., Karady, G. (1992), 'Advancement in the Application of Neural Networks for Short-Term Load Forecasting', IEEE Transactions on Power Systems, Vol.[7], pp. 250 - 257 , 1992.

[5] Pyle, D. (1999), Data Preparation for Data Mining, Morgan Kaufmann Publishers, Los Altos, CA, 1999.

[6] Mamdani, E., Assilian, S. (1975), An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller, International Journal of Man-Machine Studies, Vol. [7], No.(1), pp.1–13, 1975.

[7] Sugeno, M., Takagi, T. (1985), Fuzzy Identification of Systems and its Application to Modeling and Control, IEEE Transactions on Systems and Machine Cybernetics, Vol. [15], pp.116–132, 1985.

[8] Arafeh, L., Singh, S., Putatunda, S. (1999), 'A Neuro Fuzzy Logic Approach to Material Processing', IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. [29], No.(3), pp. 362-370, 1999.

[9] Jang, J. (1992), Neuro-Fuzzy Modeling: Architectures, Analyses and Applications, Ph.D. Thesis, University of California, Berkeley, 1992.

[10] Jang, J. (1993), ANFIS: Adaptive-Network-Based Fuzzy Inference System, IEEE Transactions on Systems and Machine Cybernetics, Vol. [23], No.(3) pp. 665–685, 1993.

[11] Rong, L., Wang, Z. (1996), An Algorithm of Extracting Fuzzy Rules Directly from Numerical Examples by Using FNN, Proceedings of the IEEE International Conference on Systems and Machine Cybernetics, Beijing, China, pp. 1067–1072, 14–17 October 1996.

[12] Ouyang, C., Lee, S. (2000), A Hybrid Algorithm for Structure Identification of Neuro-Fuzzy Modeling, Proceedings of the IEEE International Conference on Systems and Machine Cybernetics, Nashville, Tennessee, pp. 3611–3616, 8–11 October, 2000.

[13] Rojas, I., Pomares, H., Ortega, J., Prieto, A. (2000), A Self-Organized Fuzzy System Generation from Training Examples, IEEE Transactions on Fuzzy Systems, Vol. [8], 23–36, 2000.

[14] Guler, I., Ubeyli, E. (2004), Application of Adaptive Neuro-Fuzzy Inference System for Detection of Electrocardiographic Changes in Patients with Partial Epilepsy using Feature Extraction, Expert Systems with Applications, Vol. [27], pp. 323–330, 2004.

[15] Levenberg, K. (1944), A Method for the Solution of Certain Nonlinear Problems in Least Squares, Quart Appl Math , Vol. [2], pp. 164–168, 1944.
[16] Marquardt, D. (1963), An Algorithm for Least-Squares Estimation of Nonlinear Parameters, Journal of the Society for Industrial and Applied Mathematics., Vol. [11], pp. 431–441, 1963.
[17] Dennis, J. (1977), Nonlinear Least-Squares, State of the Art in Numerical Analysis, Academic Press, Orlando, FL, 1977.
[18] Jang, J., Sun, C. and Mizutani, E. (1997), Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice-Hall, Englewood Cliffs, NJ, Chapter One, Chapter Eight, and Chapter Twelve, pp. 7-9 and pp. 219-336, 1997.
[19] Werbos, P. (1974), Beyond Regression: New Tools for Prediction and analysis in the behavioral sciences, PhD Thesis, Harvard University, 1974.
[20] Rumelhart, D., Hinton, G., Williams, R. (1986), Learning Internal Representations by Error Propagation, Parallel Distributed Processing - Explorations in the Microstructure of Cognition, Vol. [1], Chapter 8, pp. 318-364, The MIT Press: Cambridge , 1986.
[21] Karayiannis, N. (1996), Hybrid Learning Schemes for Fast Training of Feed-Forward Neural Networks, Department of Electrical and Computer Engineering, University of Houston, USA, Mathematics and Computers in Simulation, Vol. [41], pp. 13–28, 1996.
[22] Hartigan, J., Wong, M. (1979), "A K-Means Clustering Algorithm," Applied Statistics, Vol. [28], pp. 100-108, 1979.
[23] Satish, B., Swarup, K., Srinivas, S., Hanumantha, A. (2004), Effect of Temperature on Short Term Load Forecasting using an Integrated ANN, Electric Power Systems Research, Vol. [72], pp. 95-101, 2004.
[24] Oriqat, Y. (2007), 'Modeling Techniques Applied to Short Essay Auto-Grading Problem', M.Sc. Thesis, Al-Quds University, pp. 46-47, 2007.
[25] Hwang, K., Kim, G. (2001), 'A Short-Term Load Forecasting Expert System', Proceedings of The Fifth Russian-Korean International Symposium on Science and Technology, pp. 112 – 116, 2001.
[26] Papadakis, S. (1998), A Novel Approach to Short-Term Load Forecasting using Fuzzy Neural Network, IEEE Transactions on Power Systems, Vol. [13], No. (2), pp. 480–492, 1998.
[27] Bhattacharyya, S., Thanh, L. (2004), Short-Term Electric Load Forecasting Using an Artificial Neural Network: Case of Northen Vietnam, International Journal of Energy Research, Vol. [28], pp. 463-472, 2004.
[28] Tamimi, M., Egbert, R. (2000), Short Term Electric Load Forecasting via Fuzzy Neural Collaboration, Electric Power Systems Research, Vol. [56], pp. 243–248, 2000.

Rae'd Basbous received the B.Sc. Degree in Computer Systems Engineering from Palestine Polytechnic University, Palestine, in 2000, and the M.S. Degree in Electronic and Computer Engineering from AlQuds University, Palestine, in 2009.

Labib Arafeh, Associate Prof. in Computing, Held his PhD from University of Manchester Institute of Science & Technology in 1992, M.Sc. from University of Essex in 1984, and B.E. from Bangalore University in 1980.

Improving Software Quality through Requirements Elicitation

Sereen Abu Aisheh
Faculty of Technology and Applied Sciences/ Al-Quds Open University, Nablus, Palestine
sabueshih@qou.edu

Abstract

Today's IT challenge is to deliver, as quickly as possible - and within a fixed budget, quality, business-critical software systems that can support business initiatives in a changing business environment, this means that three factors should be maintained to produce the desired software, these would be cost, quality and time.

Most project management efforts concentrate on meeting time and cost constraints, passing over the quality factor, research in software development industry shows that the major problem facing software development isn't crossing time and budget limits (though it's a big issue), but it's the production of software systems that wont be used as they don't address vital business needs, this is definitely a quality issue.

This paper focuses on improving the quality of a software products through improving the requirements elicitation process, it employs the results of a research conducted in two local software houses to reveal the relationship between software quality and requirements elicitation process as the road to producing better software, it also discusses two issues that may lead to failure in elicitation process (bad communication and requirements volatility).

Keywords: *Requirements Elicitation, Evolving Requirements, Software Quality.*

Introduction

IT industry still has a gap between what clients need and what they really get, most software development projects are never completed because they run out of budget and time; and even completed ones are of poor quality.

According to Chaos Report 2003 for example, clients only got 54% of the functions they requested, and 42% of the delivered functions were unused for long periods, the cause of these shortcomings was attributable to changing user requirements.

As the report indicates, there is a big gap between theory and practice in requirements area, software developers have many tools and procedures for managing client requirements and translating them

into a working software, but these tools are rarely used; as the time-to-market pressure increases, most companies tend to put less efforts in the area of requirements management, which will in turn affect the overall product quality. Statistics show that poor quality has negative effects on the long run specially for budget; the less the product quality is, the more modifications it needs, modifying the product means more time and cost, which may make the overall development project unprofitable for both client and developer.

Most developers seek the solution for this defect in the area of production tools and procedures; they focus on using some brand technologies, while the real problem lays in the requirements management process, like all other projects, software development projects need good project management process to produce a 'quality' product that meets the client's demands without crossing time and budget constraints, this means that there are predefined deliverables that should be produced after each development activity, but what if those deliverables are not what clients really want? What if they couldn't be produced?

In software development, deliverables represent the desired system functionalities that were specified based on the elicited client requirements, so if the wrong requirements were elicited; the whole project will be a failure as a result and the final product may not be used at all. To prevent such failures, more efforts should be put in order to enhance the requirements management process.

According to SQS report 2006; the problems in software development projects that failed or needed considerable additional efforts to be completed were [13]:

- Shortcomings in the specification of the requirements (50%)
- Shortcomings in the management of client requirements (40%)

The reason of failure is because most companies don't have a structured procedure for requirements management, or some do have standardized procedures but those are not implemented consistently in practice. In a research conducted for the purpose of this paper with 100 IT employees in two major IT companies in Palestine, 62% of the interviewed IT personnel stated that their

companies don't have well-defined procedures to understand clients' needs, 33% of them stated that their companies do have such procedures but those are rarely implemented in software development projects.

In addition, SQS statistics show that clients are paying additional 20% of the original contract value in average for changing requirements, our research results supports this finding as 40% of interviewed IT personnel admitted that in most cases their client asked for changes after the system was delivered, which entail more time and cost, and affected the overall quality of the product in turn.

Requirements Elicitation

Requirements elicitation is the first stage in building an understanding of the problem that the software is to solve [10]. Technically, elicitation is a process where clients, users, and developers reveal and articulate their requirements [14] but it doesn't mean that requirements are all there and can be easily captured by using any appropriate technique [11]. Most requirements management methods presume that requirements are explicitly and completely stated; however, experience shows that requirements are rarely complete and usually contain implicit requirements, software requirements characteristically suffer from inconsistency, incompleteness, ambiguity, duplication, and inconstancy [12], the way to overcome the fuzziness of requirements is by applying a structured elicitation process that deals with fact-finding, information gathering, and integration in order to obtain a set of requirements which describe the characteristics of the possible solution(s) [1].

Requirements Elicitation Problems

Problems of requirements elicitation can be grouped into three main categories [1, 12, and 14]:

- Problems of scope, in which the requirements may address too little or too much information (i.e. defining the boundary of the system).

- Problems of communication between the communities participating in the development process (e.g. users, stakeholders, and developers):

- Problems of volatility: the changing nature of requirements as they evolve over time which represents the main obstacle in elicitation process.

Our interest here is in studying the effect communication and volatility problems on the elicitation process and hence on software quality, next we discuss these problems in more details.

Problems of Volatility

Requirements change [6, 11. 14, 2]. During the time it takes to develop a system users' needs may mature because of increased knowledge brought on by the development activities, or they may shift to a new set of needs because of organizational or environmental pressures. If such changes are not accommodated, the original requirements set will become incomplete, inconsistent with the new situation, and potentially unusable because they capture information that has since become out of date.

One primary cause of requirements volatility is that user needs evolve over time. The requirements engineering process of elicit, specify, and validate should not be executed only once during system development, but rather should be returned to so that the requirements can reflect the new knowledge gained during specification, validation, and subsequent activities. Requirements management process should be iterative in nature, "so that solutions can be revised in the light of increased knowledge" [1].

Another cause of requirements volatility is that requirements are the product of the contributions of many individuals that often have conflicting needs and goals. Due to political climate and other factors, some times the needs of a particular group may be overemphasized in the elicitation of requirements. Later prioritization of the elicitation communities' needs may correct this mistake and result in requirements changes. Both the traceability of requirements and their consistency may be affected if these changes are frequent and not anticipated [1, 4].

Organizational complexity is another cause of requirements volatility as organizational goals, policies, structures, and work roles of intended end users all may change during the system's development, especially as the

number of users affected by a system's development increases.

Problem of Communication

Requirements management is a social process [1, 9] it involves various communities with different backgrounds and needs, any elicitation process that ignores this social factor will absolutely fail in understanding the characteristics of the future software.

One factor that may influence the degree of understanding is language; if clients and developers speak different languages, then the probabilities of misunderstanding what clients really want are maximized. Another factor that disrupts effective communication is the way clients express their demands, since they don't have much knowledge in computer domain so they can't articulate their needs in a form that can be understood by developers.

Problems of communication and requirements volatility proved to be critical issues as they may lead to building unsatisfactory software in the long run, our research reveals that difficulties in communication with client negatively affected the development process; 32% of respondents think that when the requirements management sessions (like JAD) were ill-structured, they had troubles understanding what their client really want, 17% of those also think that their companies didn't spend enough time and efforts in the requirements definition activity as they met their client few times only at project start. When asked about language difference between client and developer, 30% stated that when they were involved in projects for foreign clients, language difference was an obstacle as they couldn't understand client requirements and in some cases those requirements were interpreted incorrectly. 52% of respondents also stated that most clients can't speak for them selves or they don't really know what they need which in turn caused some requirements to be missing.

Respondents also said that in almost every project, clients keep changing their minds, they ask for a lot of modifications especially for functional requirements, 86% of respondents stated that the requirements statement was updated frequently due to changing client requirements, which caused the document to be inconsistent, 74% stated that the frequent modifications made it harder for

them to design the target system as they were 'lost' and ultimately their client was unsatisfied with the final product.

Obviously, research result indicate that bad communication and evolving requirements can cause the requirements management process to fail, since the actual requirements which are the base of development process can't be elicited and documented, or the wrong requirements were defined and built, in both situations the resulted software was of poor quality as 38% of respondents declared, because its either not what clients expect and need, or it didn't provide all demanded functions.

Quality is typically defined in terms of conformance to specification, freedom of defects, and fitness for purpose [3, 8]; IEEE glossary has many definitions for software quality [7]:

- The totality of features and characteristics of a software product that bear on its ability to satisfy given needs.
- The degree to which software possesses a desired combination of attributes.
- The degree to which clients or users perceive that software meets their composite expectations.
- The composite characteristics of software that determine the degree to which the software in use will meet the expectations of the client.

According to the previous definitions, a software product quality is basically measured by the degree to which the specified software accomplishes clients' expectations and desired functions, any software product that lacks those features is considered to have poor quality, this matches the findings in our research as 84% of respondents considered post delivery modifications as indication of poor product quality.

To guarantee high quality software, attention should be paid to the quality of development process itself, quality assurance procedures should be applied to monitor development activities as well as their outcomes. For any quality assurance procedure, two questions need to be investigated on a regular basis [13]:

- Is the right system being build?
- Is the system being build correctly?

These questions take us back to requirements. In order to develop a high-quality system that fulfills client needs, the right client requirements should be specified, this can't be achieved unless an interactive requirements management procedure is used to continuously refine and insure the quality of requirements list through the elicitation, specification and validation process, mainly more efforts should be put to improve requirements elicitation in order to handle the changing nature of requirements [1, 5], this requires using suitable methods for an iterative elicitation process. In our research, 50% of respondents stated that using iterative method like prototyping helped them better understand their client needs and manage the development process, as the clients witnessed the information they provided revolving into a working product, they were more excited about contributing in the development process. On the other hand, the 'knowledge diffusion' created by the prototype helped clients repair any wrong or inappropriate requirement they had provided previously or provide any missing ones. Moreover, implementing iterative elicitation process improved product quality as it minimized product modifications, especially after delivery modifications, which made the development process more profitable for both clients and development companies.

Conclusion

Software quality is a critical issue in software development; many systems failed or were evaluated to have poor quality as they don’t provide the essential business functions. The quality problem occurs due to the defect in the requirement management process, a big gap exists between theory and practice in requirements area because it's usually considered to be less important than other development activities. Problems in requirements occur either because the elicitation process is not a systematic one or because the standardized process is not performed correctly, which will cause the wrong requirements to be elicited and the wrong product to be built.

Ideal elicitation process should deal with problems of scope, communication, and requirements volatility. Implementing an iterative elicitation process on one hand can

manage the changing nature of requirements and facilitate communication between the communities participating in the development process on the other hand, which minimizes requirements errors and improves the overall software quality.

References:

[1] Christel, Michael G. and Kang, Kyo C., Issues in Requirements Elicitation, 1992.

[2] Etien, Anne and Salinesi, Camelli, Managing Requirements in a Co-Evolution Context, 2005

[3] Fitzpatrick, Ronan, O'Shea, Brendan and Smith, Peter, Software Quality Revisited.

[4] Herlea, Daniela Elena, Users' Involvement in the Requirements Engineering Process.

[5] Hickey, Ann M. and Davis, Alan M., Requirements Elicitation and Elicitation Technique Selection: A Model for Two Knowledge-Intensive Software Development Processes, 2002.

[6] Hochmüller, Elke, Quality Improvement Through Quality Requirements Management.

[7] Institute of Electrical and Electronics Engineers. IEEE Standard Glossary of Software Engineering Terminology.

[8] Lanman, Jeremy T., Software Quality – Measurements and Management, November 2001.

[9] Leite, Julio Cesar S P., A Survey on Requirements Analysis. Advanced Software Engineering Project Technical Report RTP-071, University of California at Irvine, Department of Information and Computer Science, June 1987.

[10] Mead, Nancy R., Requirements Engineering for Survivable Systems, September 2003.

[11] Nuseibeh, Bashar and Easterbrook, Steve, Requirements Engineering: A Roadmap

[12] Playle, Greg and Schroeder, Charles, Software Requirements Elicitation: Problems, Tools, and Techniques.

[13] Software quality paper: Requirements management potential and trends SQS software quality systems, March 2006.

[14] Toro, A. Duran, Jimenez, B. Bernardez, Cortes, A. Ruiz, and Bonilla, M. Toro., A Requirements Elicitation Approach Based in Templates and Patterns

Academic Researcher Information Extraction from the WEB (ARIEW)

Yousef Abuzir and Sondos kittane
Faculty of Technology and Applied Sciences/ Al-Quds Open University, Al-Bireh, RamAllah, Palestine
yabuzir@qou.edu ,sondos_kittane@hotmail.com

Abstract

Web is a large and growing collection of texts. This amount of text is becoming a valuable resource of information and knowledge. To find useful information in this source is not an easy and fast task. People, however, want to extract useful information from this largest data repository.

Academic Researcher Information Extraction from the WEB (ARIEW) is a framework for automatic collection and processing of resource related to researchers' information in the World Wide Web. ARIEW retrieves and extracts information about researchers from many servers in the Web and combines them into a single searchable database.

This paper discusses the background and objectives of ARIEW and gives an overview of its functionality and implementation of ARIEW system used to construct specialized database about researchers.

The intention is to develop the system to integrate it with other applications for Advanced Document Management. The system can be utilized in the process of automating conference organization and its usage in real world applications.

Experimental results show that our approach to researcher profiling significantly gives accurate result and performance. The methods have been applied to find related researcher. Experiments show that the accuracy of researcher finding were significantly improved by using the proposed methods.

Keywords: *Information Extraction, Knowledge discovery, Web Mining, document Management, Agent, Crawl*

1. Introduction

The enormous growth of the World Wide Web in recent years has made it important to perform resource discovery efficiently. It is often difficult to find useful information from thousands or hundreds of WebPages for a researcher or junior students. This procedure is time consuming even for sophisticate. Academic search engine, such as Google Scholar (Harzing, A. and R. van der Wal. 2008; Kloda, L. 2007), becomes an interesting and promising

topic in recent years. However most search engines return the results to users by a list and users must scan each item/webpage one by one in order to collect and re-organize these information based on users' requirements.

Information on WEB creates difficulty for filtering relevant information for decision. A researcher may know the location of such information but has to periodically access the information using direct manipulation and navigation tools. Even if the access is automated, it is still very difficult for the researcher to select relevant information. We are motivated to develop an intelligent agent to retrieve Academic Researchers' profiles on the Internet. Extracting information about academic researcher which would interest a user is difficult. Many existing agents constructed a user profile to filter and extract the user interests. In this research, an intelligent agent is developed to extract user preference and build a database to store these information for further queries.

This paper presents an academic search engine, which is developed as an efficient tool to construct researcher's profile automatically. Moreover, some searching and indexing methods, text mining and computational linguistics for underlying this problem are exploited.

1. Related Work

WEB presents a huge resource of useful unstructured information and knowledge which makes it difficult to extract and retrieve a relevant data from those sources. Therefore, there is a great necessity for information extraction (IE) systems that extract information from the Web pages and transform into program-friendly structures such as a relational database. Many approaches for data extraction from Web pages have been developed. (Chang et al., 2006) present a survey of the major Web data extraction approaches and they suggest three criteria that provide qualitatively measures to evaluate various IE approaches. These three criteria are:

- **The task domain** explains why an IE system fails to handle some Web sites of particular structures
- **The techniques used** to classify IE systems based on the techniques used.
- The third criterion is **the automation degree** for IE systems.

Information extraction is an important task with many practical applications, and many research efforts have been made so far. Many text or/and web applications like Opinion mining from

noisy text data (Dey, L. and Haque, Sk. M., 2009), Social Network Extraction of Academic Researchers (Tang J. et al 2007; Tang J. et al 2008), contact information search, question answering (Maiorano S., 2006)., integration of product information from websites (Li, L. et al 2007; Yang Z et al 2007; Yang Z et al 2010), biomedical text mining (Lourenço, 2009; Kheau, 2011), and removal of the noisy data benefit from information extraction are applications of information extraction. In these researches different methods and techniques were used. For example, rule learning based method, classification based method, and sequential labeling based method are the three state-of-the-art methods (Tang et al., 2007b). Fig 1 gives a summary of these different information extraction methods.

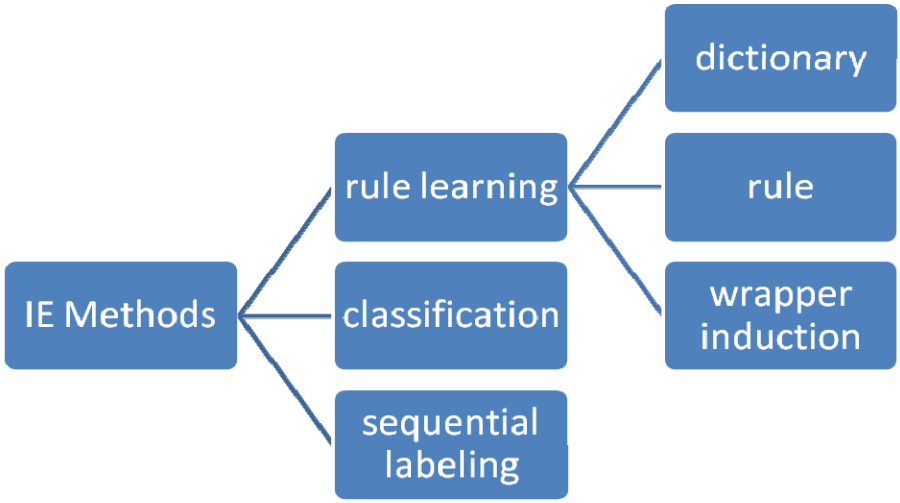


Fig 1 The Different Information Extraction Methods.

Downey (Downey D., et al 2004) introduced a Web information extraction system (KnowItAll) that uses the learned patterns as both extractors (to generate new information) and discriminators (to assess the truth of extracted information). Experimentally, by using learned patterns as extractors, they were able to boost recall by 50% to 80%; and by using such patterns as discriminators they were able to reduce classification errors by 28% to 35%.

Other work used Conditional Random Fields (CRF) model (Liu W. and Zeng J., 2011) to extract academic papers from web pages. The extensive experiments show the effectiveness of the approach. They improved the

extraction performance by exploiting the neighboring relations among the academic paper properties. Multi-Theoretical Multi-Level (MTML) (Zarandi M. F., et al, 2011) used as a framework which investigates social drivers for network formation in the communities with diverse goals. This framework serves as the theoretical basis for mapping motivations to the appropriate domain data, heuristic, and objective functions for the personalized expert recommendation.

Another approach based on extracting contextualized user profiles in an enterprise resource sharing platform according to the users' different topics of interest was presented by Schirru (Schirru,R, et al, 2010). Each topic is represented as a weighted term vector.

Extraction Prioritization proposed as automatic technique for obtaining the most valuable extraction results as early as possible in the extraction process (Huang J. and Yu C., 2010). They formally defined a metric for measuring the quality of extraction results, which is suitable for the web retrieval context and developed statistical methods to efficiently estimate the page utilities without launching full-scale extractions.

2. Problem Analysis

Researcher Academic Profile is an important topic in research community. An academic can have different types of information: contact information (including address, email, telephone, and fax number), Academic profile (including homepage, position, portrait, affiliation, research interest, publications, and documents), and social network information (including person or professional relationships between persons, e.g. friend relationship). Figure 2 shows a sample presentation of these information. However, the information is usually hidden in heterogeneous and distributed web pages.

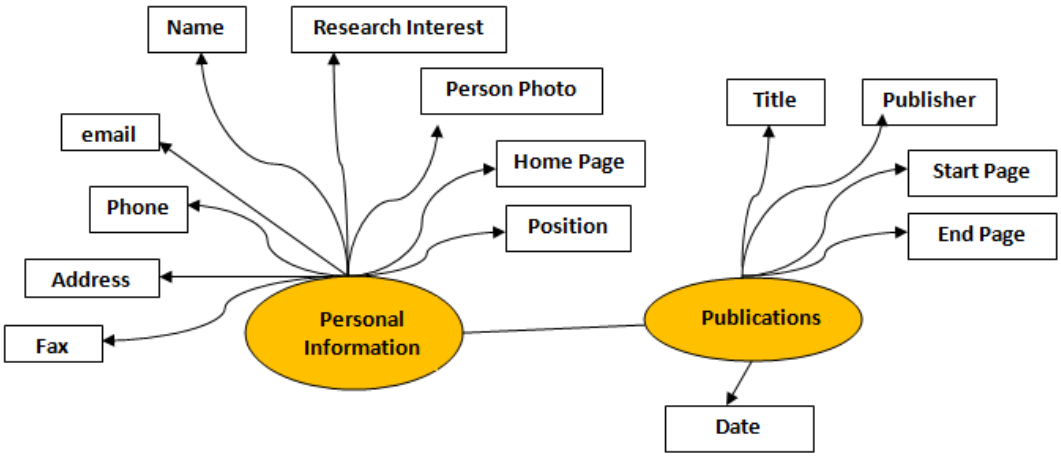


Fig 2 A Sample Presentation of the Researcher Profile's Schema

Many previous studies have indicated that a researcher's personal Web site can be considered as identity and self-presentation of the researcher on the Web, and it can be used to different and shows interesting information about the researcher (Doring, 2002). As we have discussed, a researcher ' s Web site usually has information about her/his research interest, publications, research projects, etc., which well represents this researcher.

We have investigated the problem of contact information extraction and academic profile extraction. We have found that the academic information is mainly hidden in person homepage, person introduction page (web page that introduces the person), person list (e.g. a faculty list), and email message (e.g. in signature). We employed the classification based method to extract the person information from the different types of web pages.

The usefulness of a research profile constructor system depends to a large extent on its ability to automatically determine one or more researchers profiles related to the work of interest. Various approaches exist to determine the degree of similarity of related in order to identify related work (Sabbah T. S., et al, 2009).

3. System Architecture

This section describes the architecture of our system which is an Internet agent that gathers information from the Web Pages in order to build a local database of researchers. As an application, we showed a case of an agent building a database with information about academic contacts (phone, email, Postal Address), photos, their interest's researches and publications.

Although, this system used to reduce the effort of the researcher in finding information about other researchers, the system may be used in other applications like indexing and classification, conference management. The database created by the agent was implemented using MySQL.

The system architecture in Figure 3 shows that there are five modules: the module of Concept Crawler, which is responsible for collecting data from WEB; Database module; the module of Researcher Agent; Query Processor and UI module. In the following paragraphs, these components and underlying methods are described one by one.

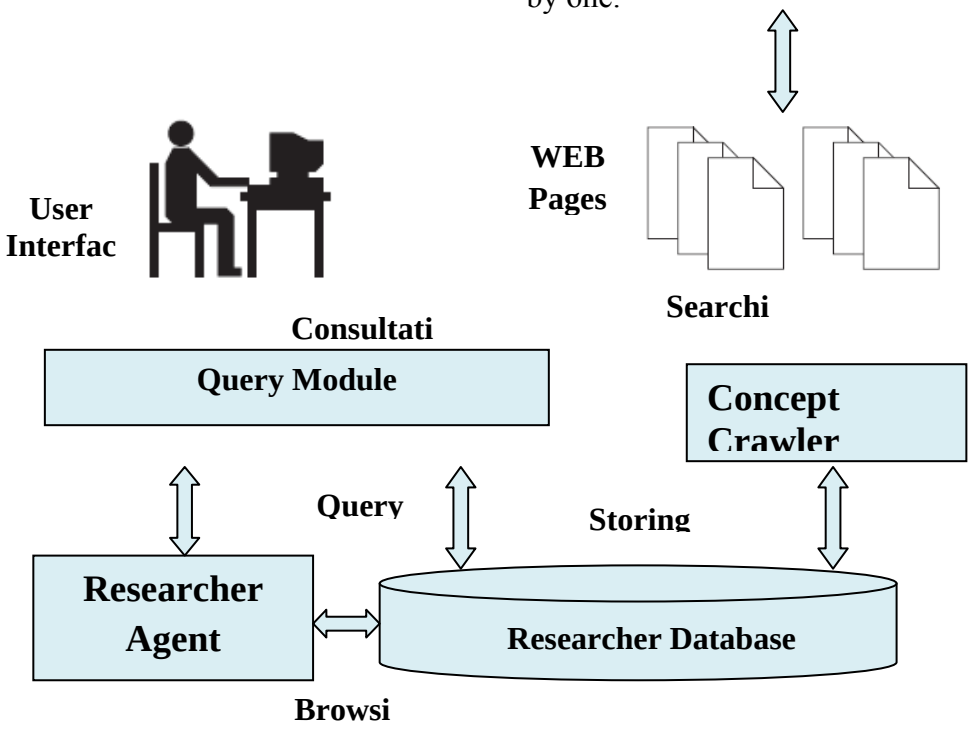


Fig. 3 ARIEW System Architecture

The module of Concept Crawler - Our system highly depends on the topic related data using Concept hierarchy, hence we use topic focused crawler to collect data. This structure is presented to evaluate WEB Pages about the specific topic. Only the WEB Pages whose score is greater than a given threshold is fetched. Two factors contribute to the score. The first one is the content of given web pages, including title, keyword, text and description. The second one is the predefined patterns in the web page. For those satisfied Web pages, we access them, analyze the content inside, organize them with XML format, and store them into data storage. Figure 2 shows the hierarchical structure of the data collecting module and the procedure for data parsing.

Database Module - This module store the result of web pages collected from the Concept Crawler and the researcher information extracted from collected web pages using the Agent to index and queries the databases. The database created by the agent was implemented using MySQL. The key contact information in the database consists of researcher name (Figure 4). The database also stores general information about the researchers includes title, photos, address, phone, email, interests, information about activities and publications.

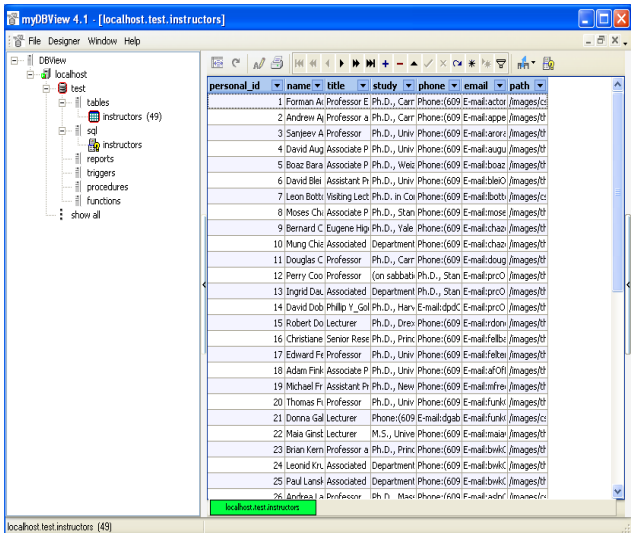


Figure 4 MySQL Database shows a sample of the extract information from select WEB Page

The Module of the Agent parses the web pages in the database to get related information from the storage component Figure 4. We use predefined patterns to find the related information about researchers. Each

keyword (except for the stop words) extracted by the module using these patterns will be an attribute of the researcher, and stored in the database as a result of indexing. After analyzing all the related web pages, agent module returns the required information about the academic researchers. Our agent module use pattern based extraction mechanisms to extract information on researcher contacts.

Based on the home pages for a researcher in the database, the module of the agent starts to extract information from these web pages and referenced pages or URL Links. Searching WEB pages is done in two different approaches. The first approach based on Keyword and called keyword-based and the second one is called pattern-based search (Figure 5). In the first approach, keyword-based search, the agent searches for keywords as specified in the extraction profile. For each keyword, a set of options is specified which tells the agent what information may be found in proximity to the keyword. Although such keyword searching is relatively simple, it has proved effective and is used in our system to find general information about the researchers and publication lists or project descriptions.

Our agent model use pattern based extraction mechanisms to extract information on researcher contacts. However, the agent itself generates these patterns based on the structure of individual items found in repeating items such as HTML lists and tables.

```
<--parse the Researcher objects--!>

 <"var-def name="instruct_objects>
xpath                                    >
        expression="data(//td[@class='people
                                  <"(['people_center
                          <html-to-xml>
http                                     >
 url="http://www.cs.princeton.edu/people/f
                                        </"aculty
                      <html-to-xml/>
                              <xpath/>
                               <var-def/>
```

Figure 5 A sample XPath code used by the crawl to select WEB Page

```
<-- parse the Researcher photo--!>
                  <"var-def name="photourl_objects>
xpath                                              >
 expression="//td[@class='people']//img[1]/@src
                                                  <"
                                  <html-to-xml>
http                                               >
url="http://www.cs.princeton.edu/people/faculty
                                                 </"
                                <html-to-xml/>
                                       <xpath/>
                                      <var-def/>
```

Figure 6 A sample XPath code used by the crawl to extract Photos from the WEB Page

Query Processor Module - The main task of query processor is to execute query and provide the information related to the query as a results to the user.

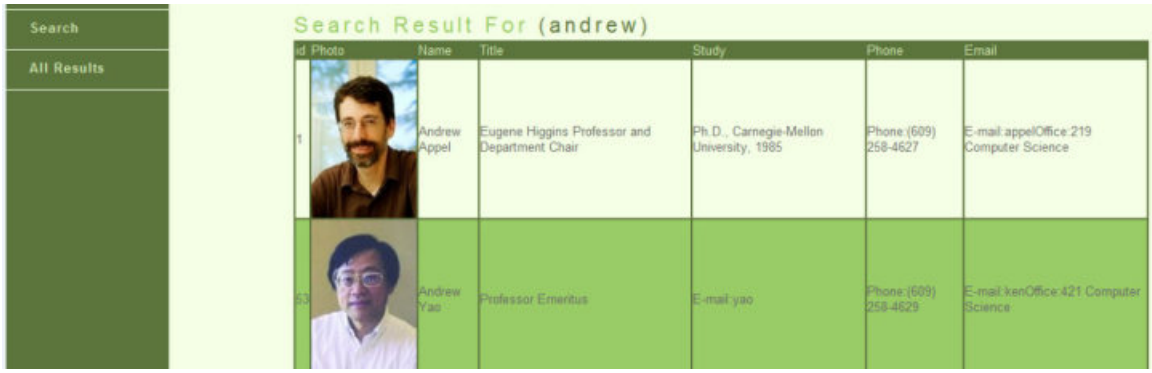


Figure 7 Results show information about the researchers in query

User Interface - A friendly browser-based user interface is presented to the end users. After submitting query keywords, the user will get a comprehensive result shown in Figure 7 which is composed of the information about the researchers in query. And by clicking each of the labels of clustering result, users can get some analysis for each sub-topic respectively, the topics are clustered hierarchically.

In addition, if a user is interested in a particular author, the system provides different information related to the author, likes name, address, email, phone, scientific interests, etc. And the user can also get the answers for the most related information too. The user accesses the database directly or retrieves and process information on researcher contact.

4. Summary

Intelligent Agent aims to facilitate the construction of researchers' profiles by decreasing the amount of effort required to construct researcher databases in special domain. However, there are few studies that attempt to automate the entire construction process from the collection of domain-specific literature, to text mining to build new database or enrich existing ones. In this paper, we present a complete framework for an intelligent Agent that enables us to retrieve documents from the Web using Concept Hierarchy crawling that identify domain-specific documents, and then perform text mining in order to extract useful information form university web pages. We have carried out several experiments on components of this framework in a

computer science domain. Other domain can be easily used by adding the concept hierarchy for that domain. This paper reports on the overall system architecture and our initial experiments on information extraction using text mining techniques to enrich the domain researcher database.

This paper presents an academic search engine, which is developed as an efficient tool to construct researcher's profile automatically. Moreover, some searching and indexing methods for underlying XML data are exploited. The paper describes the architecture and main features of the system. It also briefly presents the experimental results of the proposed methods.

Table 1 shows result of performance of the system for the stage of information extraction from different web pages. It is clear that the overall result of the pattern approach is more accurate than the keyword approach. The overall precision of the system was 84.90, which is a good indication about the performance of the system. The result differences refer to the web page and the structure of the web page.

5. Conclusion and Future Work

This paper focuses on approaches to extract valuable information from large quantities of unstructured textual information, combining methods from several research areas, including information retrieval, text mining, computational linguistics, and machine learning.

This agent is tested with different experiments. These experiments aim at examining the effect of User Profile and Concept Hierarchy used by the module of Concept Crawler, The overall precision was 91.23 for selecting a related web pages and finally, Precision of the retrieval performance (table 1).

The result shows that agent collects general interests of users when extracting the user profile of the academic Researcher's profile from the WEB. This research indicates agent with using both concept hierarchy and user profile approaches can achieve good result in information retrieval performance.

As future work, such an agent can provide a value added service by using information extracted from Web documents to maintain the database and ensure its currency. The agent may either update the database directly

or consult with the user as to whether or not it should perform the updates.

TABLE 1

PERFORMANCE OF THE SYSTEM BASED ON PATTERN AND KEY WORD APPROACH.

Researcher Profile Information	Patterns Approach		Keyword Approach	
	Precision	Recall	Precision	Recall
Name	91.09	89.32	87.98	89.99
Photo	90.32	88.41	73.12	58.87
Phone	89.75	91.89	76.95	83.25
email	83.21	84.28	81.77	78.32
fax	92.54	89.78	73.12	75.45
address	87.90	84.89	77.98	80.24
Interesting topics	66.78	64.45	59.23	63.47
Position	77.57	65.01	73.99	57.67
Result	**84.90**	**82.26**	**75.52**	**73.40**

References

[1] Harzing, A. and R. van der Wal. (2008). "Google Scholar as a new source for citation analysis." Ethics in Science and Environmental Politics (ESEP) 8(1):61–73. doi:10.3354/esep00076

[2] Kloda, L. (2007). "Use Google Scholar, Scopus and Web of Science for comprehensive citation tracking." Evidence Based Library and Information Practice2(3):87.

[3] Chang, C.-H.; Kayed, M.; Girgis, R.; Shaalan, K.F.; (2006), IEEE Transactions on Knowledge and Data Engineering, (2006), 18(10), pp 1411 – 1428, DOI: 10.1109/TKDE.2006.152

[4] Dey, L. and Haque, Sk. M., (2009), Opinion mining from noisy text data, International Journal on Document Analysis and Recognition, Volume 12, Number 3, pp 205-226. Doi: 10.1007/s10032-009-0090-z.

[5] Tang J., Zhang D, and Yao L.,(2007a) Social network extraction of academic researchers. In ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp 292–301,

[6] Tang J., Zhang J., Yao L, and Li J., (2008). Extraction and mining of an academic social network. In WWW '08: Proceeding of the 17th international conference on World Wide Web, pp 1193-1194, New York, NY, USA, 2008. ACM.

[7] Maiorano S., (2006), Question answering: Technology for intelligence analysis. In Tomek Strzalkowski and Sanda Harabagiu, editors, Advances in Open Domain Question Answering, volume 32 of Text, Speech and Language Technology, chapter 16, pages 477–504. Springer Netherlands, Dordrecht, 2006.

[8] Li L., Liu Y., Obregon A., Weatherston M., (2007), Visual Segmentation-Based Data Record Extraction from Web Documents, Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on In Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on (2007), pp. 502-507. doi:10.1109/IRI.2007. doi:10.1016/j.jbi.2009.04.002

[9] Yang Z., Li L., Wang B.and Kitsuregawa M., (2007) Towards Efficient Dominant Relationship Exploration of the Product Items on the Web. AAAI 2007, pp1483-1488

[10] Yang Z., Li L., Wang B.and Kitsuregawa M., (2010) Efficient Analyzing General Dominant Relationship Based on Partial Order Models. IEICE Transactions 93-D(6): 1394-1402 (2010)

[11] Lourenço A., Carreira R., Carneiro S., Maia P, Glez-Peña D., Fdez-Riverola F., Ferreira E. C., Rocha I., and Rocha M.(2009), @Note: a workbench for biomedical text mining. Journal of biomedical informatics, 42(4):710–720, August 2009.

[12] Kheau, C. S., Alfred, R. and Obit, J. H., (2011), BioDARA: Data Summarization Approach to Extracting Bio-Medical Structuring Information, Journal of

Computer Science 7(12), pp1914-1920, doi: 10.3844/jcssp.2011.1914.1920

[13] Tang J., Hong M., Zhang D., Liang B., and Li J.,(2007b). Information Extraction: Methodologies and Applications. In the book of Emerging Technologies of Text Mining: Techniques and Applications, Hercules A. Prado and Edilson Ferneda (Ed.), Idea Group Inc., Hershey, USA, 2007. pp. 1-33

[14] Downey D., Etzioni O., Weld D. S., and Soderland S. (2004),. Learning Text Patterns for Web Information Extraction and Assessment. Proceedings of the AAAI-04 Workshop on Adaptive Text Extraction and Mining, 2004.

[15] Liu W. and Zeng J.,(2011) Automatically Extracting Academic Papers from Web Pages Using Conditional Random Fields Model, JOURNAL OF SOFTWARE, VOL. 6, NO. 8, AUGUST 2011, pp1409-1416

[16] Zarandi M. F., Devlin H. J., Huang Y., and Contractor N.,(2011),. Expert recommendation based on social drivers, social network analysis, and semantic data representation. In Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '11). ACM, New York, NY, USA, 41-48. DOI=10.1145/2039320.2039326.

[17] Schirru R., Baumann S, Memmel M and Dengel A (2010), Extraction of Contextualized User Interest Profiles in Social Sharing Platforms, Journal of Universal Computer Science, vol. 16, no. 16 (2010), pp 2196-2213.

[18] Huang J. and Yu C., (2010). Prioritization of Domain-Specific Web Information Extraction. In Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010), AI & the Web Special Track. Atlanta, GA, USA. July, 2010.

[19] Doring, N. (2002). Personal Home Pages on the Web: A Review of Research, Journal of Computer Mediated Communication, 7(3), pp 1-28.

[20] Sabbah T. S., Jayousi R. and Abuzir Y. (2009),"Schema Matching Using Thesaurus", The 3rd Int. Conference on Software, Knowledge Management and Applications, SKIMA 2009, Fez, Moroco,2009.

A Comparative Study of Statistical and Data Mining Algorithms for Prediction Performance

Amjad Harb and Rashid Jayousi
Al-Quds University
Jerusalem, Palestine
amjad@pcbs.gov.ps and rjayousi@science.alquds.edu

Abstract

The aim of this study is to perform a comparison experiment between statistical and data mining modelling techniques. These techniques are statistical Logistic Regression, data mining Decision Tree and data mining Neural Network. The comparison will evaluate the performance of these prediction techniques in terms of measuring the overall prediction accuracy percentage agreement for each technique. The ratio of the binary values of the dependent variable in the training dataset and the population is used on the three techniques to find the effect of this ratio on the prediction performance. For a given data set, the results shows that the performance of the three techniques is comparable in general with small outperformance for the Neural Network. An affecting factor that makes the prediction accuracy varied is the dependent variable values distribution (distribution of "0"s and "1"s). It is seen that, for all of the three techniques, the overall prediction accuracy percentage agreement is high when the ratio of "0"s and "1"s is 3:1, whereas for the ratios 2:1 and 1:1 the performance is lower.

Keywords**:** *Data Mining, Classification, Prediction Model, Statistical Logistic Regression, Neural Network, Decision Tree.*

1. Introduction

An important and challenging area of research is information management. Historical data was analyzed using several ways for hidden knowledge extraction that can help in decision making. This is called Knowledge Discovery or Data Mining. The popular goal from data mining is prediction and the popular data mining technique used for prediction is classification. Classification can be accomplished statistically or by data mining methods. [1]

Prediction techniques performance comparison issues is an interesting topic for many researchers. A comparative study by Lahiri R. [1] compared the performance of three statistical and data mining techniques on Motor Vehicle Traffic Crash dataset, resulted that the data information content and dependent attribute distribution is the most affecting factor in prediction performance. Delen D. et al. [2] targeted data mining methods comparison as a second objective in the study, while the main objective was to build the most accurate prediction model in a critical field, breast cancer survivability. In the same area, Artificial Intelligence in Medicine, Bellaachia A. et al. [3] continued the

work of [2] and improved the research tools especially the dataset. An important application area that exploited data mining techniques heavily was the network security. Panda M. et al. [4] also performed a comparative study to identify the best data mining technique in predicting network attacks and intrusion detection. Also the data contents and characteristics revealed as an affecting factor on the data mining and prediction algorithms performance.

In this research we will continue on the work of Lahiri R. [1] to perform a comparison on the same data mining techniques: Logistic Regression, Neural Network and Decision Tree, but with more accurate data content and quality. Also we will work on Lahiri's future work recommendation by determining more precise predictors that significantly define and affect the output. In another words, we intended to find the effect of the most correlated variables, as predictors, to the dependent variable on the prediction accuracy rates for the aforementioned prediction techniques. The overall prediction percentage agreement will be the main performance metric which will be measured under different dependent variable values distributions (distribution of "0"s and "1"s) in the dataset. This is to identify the effect of the dependent variable values distribution on the overall prediction accuracy for the three prediction techniques. The experiment will exploits a historical dataset about the Palestinian Labor Force. The dependent attribute will be the individual's "Labor Force Status", that have values: 0 as Employed and 1 as Unemployed. The source of such data is the Palestinian Central Bureau of Statistics (PCBS).

This paper is organized as follows: The literature and related work will be discussed in Section 2. The research methodology to perform the experiment will be presented in Section 3. Experimental results are presented and discussed in Section 4. Finally, Conclusion is given in the last Section 5.

2. Related Work

Many studies have been done across countries on data mining. Applications of data mining were used in a large number of fields, especially for business and medical purposes.

As data mining is a new technology field, it is important and very helpful in predicting and detecting underlying patterns from large volumes of data, many researches were published, comparing results of data mining algorithms in several areas. A research by Rochana Lahiri (2006) performed a performance comparison of several data mining and statistical techniques for classification model. She used a database from Louisiana Motor Vehicle Traffic Crash. The performance was measured in terms of the classification agreement %. The effect of Decision Tree, Neural Network, and Logistic Regression models for different sample sizes and sampling methods on three sets of data had been investigated. The study concluded that a very large training dataset is not required to train a

Decision Tree or a Neural Network model or even for Logistic Regression models to obtain high classification accuracy and the overall performance reached a steady value at the sample size of 1000, irrespective of the total population size. The information content of a training dataset is an important factor influencing classification accuracy and is not governed by the size of the dataset. Another important result was that the sampling method has not affected the classification accuracy of the models. She also stated that the overall classification accuracy of the all three methods were very much comparable and no one method over performed any other. She tried to find the effect of the "0"s and "1"s distribution of dependent variable values in the dataset but because the data was very skewed, she failed to do this. As a future work, the study recommends to apply the same study on a dataset were the relationships between the dependent variable and the independent variables are more rigid. i.e.: to select predictors that strongly describe the dependent attribute, and to study the effect of "0"s and "1"s dependent variable values distribution. [1]

The data mining methods comparison were targeted as a second objective in some studies that mainly aimed to develop a prediction model in a critical fields, like medicine, by investigating several data mining methods, intending to get the model that have the highest prediction accuracy. This type of studies has been addressed by Delen D. et al. (2005) in the context of predicting breast cancer survivability. Multiple prediction models, using Artificial Neural Networks, Decision Trees, and Logistic Regression, for breast cancer survivability using a large dataset had been developed. The comparison among the three models had been conducted depending on measuring three prediction performance metrics: classification accuracy, sensitivity and specificity. The k-Fold cross-validation test was used to minimize the bias associated with the random sampling of the training and missing data. The results of the study showed that the Decision Tree (C5) preformed the best of the three models evaluated. Sensitivity analysis, which provides information about the relative importance of the input variables in predicting the output field, was applied on Neural Network models and provided them with the prioritized importance of the prediction factors used in the study. [2]

Another related study in medicine by Bellaachia A. et al. (2006) also in the context of predicting breast cancer survivability. The researchers took the study of Delen D. et al. [2] as the starting point with the same dataset source but with a newer version and different set of data mining techniques. For modeling and comparison, three data mining techniques had been investigated: the Naïve Bayes, the back-propagated Neural Network, and the C4.5 Decision Tree algorithms. The main goal was to have a prediction model with high prediction accuracy, besides high precision and recall metrics for patients' data retrieval. They

used other performance metrics: specificity and sensitivity to compare the prediction models. The results presented that C4.5 algorithm has a much better performance than the other two techniques. The obtained results differed from the study of Delen D. et al. [2] due to the facts that they used a newer database (2000 vs. 2002), a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (industrial grade tools vs. Weka). [3]

In network security, data mining techniques used heavily in predicting network intrusion detection systems to protect computing resources against unauthorized access. Several studies were performed in this area and some of them addressed the prediction performance comparison of different data mining techniques like the study by Panda M. et al. (2008). A dataset of 10% KDDCup'99 intrusion detection has been generated and used in the experiment. Three popular data mining algorithms had been used in the experiment: Decision Trees ID3, J48 and Naïve Bayes. The prediction performance metrics used in the study were the time taken to build the model and the prediction error rate. For the evaluation of prediction error rate, the 10-fold cross validation test was used. As a result of the experiment, the Decision Trees had proven their efficiency in both generalization and detection of new attacks more than the Naïve Bayes. But this maybe dependence on the contents and characteristics of the data which allows single algorithm to outperform others. [4]

Amooee G. et al. (2011) used data mining techniques to identify defective parts manufactured in an industrial factory and to maintain high quality products. A data of 1000 records was collected from the factory and 10% (100 records) of the data was about a defective parts. Prediction accuracy and processing time of the prediction techniques were the comparison performance metrics. The results showed that SVM and Logistic regression prediction algorithms has the best processing time with high overall prediction accuracy. The decision tree with its tree different branching algorithms (CRT, CHAID, and QUEST) achieved the highest prediction accuracy rates but needed more time. Neural network achieved the least prediction accuracy rate with medium processing time. [5]

Data mining concept was the most appropriate to the study of student retention from sophomore to junior year than the classical statistical methods. This was one main objective of the study addressed by Ho Yu C. et al. (2010) in addition to another objective that identifying the most affecting predictors in a dataset. The statistical and data mining methods used were classification tree, multivariate adaptive regression splines (MARS), and neural network. The results showed that transferred hours, residency, and ethnicity are crucial factors to retention, which differs from previous studies that found high school GPA to be the most crucial contributor to retention. In Ho

Yu C. et al. research, the neural network outperformed the other two techniques. [6]

The prediction techniques RIPPER, decision tree, neural networks and support vector machine were used to predict cardiovascular disease patients. The performance comparison metrics were the Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Kumari M. et al. study showed that support vector machine model outperforms the other models for predicting cardiovascular disease. [7]

The neural network was found to achieve better performance compared to the performance rates of Naive Bayes, K-NN, and decision tree prediction techniques in a study performed by Shailesh K R et. al. (2011) to predict the inpatient hospital length of stay in a super specialty hospital. [8]

The same result was seen that the neural network outperformed both the decision tree and linear regression models when the performance for the students' academic performance in the undergraduate degree program was measured by predicting the final cumulative grade point average (CGPA) of the students upon graduation. The correlation coefficient analysis was used to identify the relationship of the independent variables with the predictors. Ibrahim Z. et al. (2007) [9]

Social network data, using data mining techniques and the prediction error rates were the comparison metric, was studied by Nancy P. et al. (2011). The tree based algorithms such as RndTree, ID3, C-RT, CS-CRT, C4.5, CS-MC4 and the k-nearest neighbor (k-NN) algorithms were used in the study. The RndTree algorithm achieved least error rate and outperformes the other algorithms. [10]

C. Deepa et al. (2011) compared the prediction accuracy and error rates for the compressive strength of high performance concrete using MLP neural network, Rnd tree models and CRT regression. The results showed that neural network and Rnd tree achieved the higher prediction accuracy rates and Rnd tree outperforms neural network regarding prediction error rates. [11]

The Rand tree algorithm also outperforms the other algorithms, C4.5, C-RT, CS-MC4, decision list, ID3 and naïve bayes, in a study of vehicle collision patterns in road accidents by S.Shanthi et al. (2011). Selection algorithms were used including CFS, FCBF, Feature Ranking, MIFS and MODTree, to improve the prediction accuracy. Feature Ranking algorithm was found the best in improving the prediction accuracy for all algorithms. [12]

In this study, we have used a Labor Force Data (LFS 2009) in the Palestinian's territories as a training dataset with 38,037 records. For testing, we used the same dataset of Labor Force Data (LFS 2009) and Labor Force Data (LFS 2010). The two datasets was processed and cleaned against missing values and inconsistency.

3. Methodology

To achieve the objectives of this research, we have started data preparation (LFS 2009 and LFS 2010)

in a suitable form for the experiment. The data contains the following variables: person's labor status (dependent variable), Sex, Age at last Birthday, Does he currently attending school?, Years of schooling, Educational Attainment (higher qualification), Refugee Status, District, Locality Type, Region, and Marital Status. All these variables are numeric data type and nominal measure, see Table 1. We should mention that for the training of the prediction algorithms we used LFS 2009 dataset and for testing we used both LFS 2009 and LFS 2010.

TABLE 1: SPECIFICATION OF THE LABOR FORCE DATASET VARIABLES

Variable	Values	Measure
Sex	1 or 2	Nominal
Age at last Birthday	0-100	Nominal
Does he currently attending school	1-4	Nominal
Years of schooling	0-40	Nominal
Educational Attainment (higher qualification)	1-10	Nominal
Refugee Status	1-3	Nominal
District	16 Category	Nominal
Locality Type	1-3	Nominal
Labor Force Status	0: Employed 1: Unemployed	Nominal
Region	1 or 2	Nominal
Marital Status	1-3	Nominal

We selected the "*person's labor status"* as the dependent variable that has only two expected values, 1 as Unemployed and 0 as Employed. Because we need to find the effect of dependent variable values distributions (distribution of "0"s and "1"s) on the overall prediction percentage agreement, we prepared different copies of the dataset and changed the dependent variable values distributions for three values ratios, see Table 2.

TABLE 2: DEPENDENT VARIABLE VALUES DISTRIBUTION

Ratio	0: Employed	1: Unemployed	Total
1:1	9,292	9,292	**1,8584**
2:1	20,122	9,292	**29,414**
3:1	28,745	9,292	**38,037**

The original dataset is the set with ratio 3:1, and from this dataset we extract the other two datasets by reducing the number of records of the "Employed" persons in a way to have the required ratio. This was done by deriving a stratified sample for each ratio type using "District" as stratifying variable.

Selecting the independent variables (inputs) that have tight relationships to the dependent variable (output) to get more accurate results when applying the aforementioned prediction techniques, is one of the main objectives of this study. To verify this, we applied *"Spearman's Bi-variate correlation"* analysis to identify the exact coefficient percentage and significance of this effect between the variables of the dataset on each other. The correlation analysis was performed on the three ratio datasets and only the correlation

results of the dependent variable *"Labor Force Status"* with other variables was selected, see Table 3. As shown in Table 3, we selected the variables that have correlation coefficient of 10% and more, and significance value of 0.05 and less. This results the variables, marked in bold, that are the most related to the dependent variable, they are (ordered by importance): Region, Age, District, Marital Status, and Refugee Status.

After this step, we run the prediction techniques: Binary Logistic Regression, Multilayer Perceptron Neural Network (MLP), and CRT Decision Tree on the three ratio datasets at two stages. In the first stage we assigned the all variables as independent variables, apply the prediction algorithm and recording the results. The same procedure was followed in stage two but we assigned the independent variables depending on the correlation analysis results as shown in Table 2.

TABLE 3: LABOR FORCE STATUS CORRELATIONS

N		1,8584	29,414	38,037
Unemployment : Employment Ratio		**1:1**	**2:1**	**3:1**
Sex	Coefficient	.001	.004	.001
	Significance	.866	.457	.790
Age at last Birthday	**Coefficient**	**-.242**	**-.219**	**-.203**
	Significance	**.000**	**.000**	**.000**
Does he currently attending school	Coefficient	.017	.014	.011
	Significance	.023	.015	.026
Years of schooling	Coefficient	.017	.019	.016
	Significance	.023	.001	.002
Educational Attainment (higher qualification)	Coefficient	-.015	-.019	-.018
	Significance	.035	.001	.001
Refugee Status	**Coefficient**	**-.117**	**-.107**	**-.099**
	Significance	**.000**	**.000**	**.000**
District	**Coefficient**	**.238**	**.221**	**.205**
	Significance	**.000**	**.000**	**.000**
Locality Type	Coefficient	-.008	-.003	-.003
	Significance	.267	.622	.539

TABLE 3 CONT.

Region	**Coefficient**	**.252**	**.243**	**.228**
	Significance	**.000**	**.000**	**.000**
Marital Status	**Coefficient**	**-.197**	**-.185**	**-.173**
	Significance	**.000**	**.000**	**.000**

The primary results obtained of the overall prediction accuracy percentage for the three prediction algorithms showed that the differences between assigning the independent variables by selecting all variables in stage one and selecting only the correlated variables in stage two are almost close. Even it is slightly larger by selecting the all variables than selecting the correlated variables. After the training of the models, we tested them using the data of LFS 2009 and LFS 2010.

The *PASW Statistics (SPSS Release 18.0.0)* from IBM was used in all operations and to calculate all the aforementioned statistical and data mining techniques and methods.

4. Experimental Results

The results of the analyses performed on the three different dataset's ratios using the three different prediction techniques have been trained and tested using the LFS 2009 and LFS 2010 datasets. Our concentration will be on the overall prediction accuracy as a metric of performance of the prediction techniques. As we have said in the

methodology, we experimented with the three prediction techniques for each ratio dataset two times. One time using the all variables as independent variables and the other time assigning only the variables that are the most correlated with the dependent variable, *"Labor Force Status"*, as seen in Table 2. The overall prediction accuracy for the two iterations in each ratio dataset are too comparable, even with using the all variables as independent variables resulted a bit larger accuracy than using the correlated variables and for all ratio datasets, see Table 4. This results suggest that the three prediction techniques, in some way or another, rely on the independent variables that are most correlated to the dependent variable and with more higher accuracy if some of the other less correlated independent variables, that also have acceptable significance value, added to the analysis. We can imply from this that we can save time by selecting all candidate variables, by common sense, as independent variables without worry to calculate the correlations, unless we can find another way to identify the most correlated variables. Depending on this, we continued the analysis without the results that were based on the correlated variables.

TABLE 4: OVERALL PREDICTION ACCURACY PERCENTAGE OF THE PREDICTION TECHNIQUES FOR BOTH *ALL* AND *CORRELATED* VARIABLES

Method*	Type	Ratio	Training	Testing 2009	Testing 2010
DT	Corr.	1:1	65.7	67.1	65.3
DT	All	1:1	68.5	67.2	68.4
DT	Corr.	2:1	72.5	77	78.3
DT	All	2:1	73.4	77.1	76.8
DT	Corr.	3:1	77.4	77.4	79
DT	All	3:1	78.1	78.1	78.8
LR	Corr.	1:1	64.2	62.4	62.7
LR	All	1:1	64.6	64.1	66
LR	Corr.	2:1	72.2	76.9	78.6
LR	All	2:1	72.3	77	78.2
LR	Corr.	3:1	77.5	77.5	79
LR	All	3:1	77.6	77.6	78.7
NN	Corr.	1:1	67.5	67.3	65.9
NN	All	1:1	70.5	70.5	70.4
NN	Corr.	2:1	72.6	77	78.3
NN	All	2:1	74.9	78.3	77.6
NN	Corr.	3:1	77.6	77.6	79.1
NN	All	3:1	78.8	78.8	78.4

*: DT: Decision Tree, LR: Logistic Regression, NN: Neural Network

For investigating the effect of dependent variable values distribution we replicated the dataset into three categories each with different "0" to "1" ratio as in Table 2 in the previous section.

Fig.1 plots three graphs one for each dataset type as training, testing 2009 and testing 2010 respectively. Each graph plots the overall prediction accuracy percent over the three datasets ratios and using the prediction methods.

It is seen that, in the three graphs, the prediction performance for the three prediction techniques were comparable and no one prediction technique outperform the other two, except a very small outperformance for the neural network in some conditions. For the training graph, it is seen that when the dependent variable values distribution ratio was 1:1, the overall prediction accuracy rate of neural network, decision tree and logistic regression were around 70%, 68% and 66% respectively. This means that neural network predicted the person's work status accurately more than the other two with small difference among them. When the dependent variable values distribution ratio was 2:1, the same reading was achieved but with more smaller differences among the three prediction techniques' overall prediction performance rates. The behavior was the same when the dependent variable values distribution ratio became 3:1 but this time the overall prediction accuracy rates for the three methods were almost the same with a non significant difference among their results reached a maximum value of 1.2%. It is also seen that by increasing the dependent variable values ratio, the overall prediction accuracy rate for each one of the three techniques plot an increasing curve, starting from 1:1 ratio that scored the lowest overall prediction accuracy rate value with significant difference between this value and the prediction value when the ratio was 2:1. The same behavior was seen between the ratios 2:1 and 3:1.

For testing data the behavior was the same as for the training data. Each prediction technique plots an increasing curve. As the dependent variable values distribution ratio increased, the overall prediction accuracy rate also increased. The difference between the training data curves and the testing data curves was that in the testing curves a plateau was reached when the dependent variable values distribution ratio was 2:1 and increasing the ratio to 3:1 didn't significantly improve the overall prediction accuracy rate more. While the curves of the training data were almost linear and no plateau was reached. This means that in the testing graphs the curves met when the dependent variable values distribution ratio was 2:1 and continued the same level reaching to ratio 3:1, while in the training curves they met when the dependent variable values distribution was 3:1 and not before.

Another exciting results were the results of prediction accuracy rates for the individual values of the dependent variable. Table 5 shows the prediction accuracy rates of the dependent variable values and the overall prediction accuracy for the three prediction techniques over the three dependent variable values distribution ratios and for the testing datasets 2009 and 2010.

It is seen that for all of the prediction techniques and when the dependent variable values distribution were 2:1 and 3:1, the prediction accuracy rates for the "0" and "1" were highly not comparable. The prediction methods predicted the dependent variable value that have the largest frequency in the dataset with high accuracy rate while they fail to predict the other value at approximately the same, or nearby, accuracy rate. For example the neural network in the year 2009 data and for the ratio 3:1, predicted the value of "0" and "1" for 94.9% and 28.9% respectively.

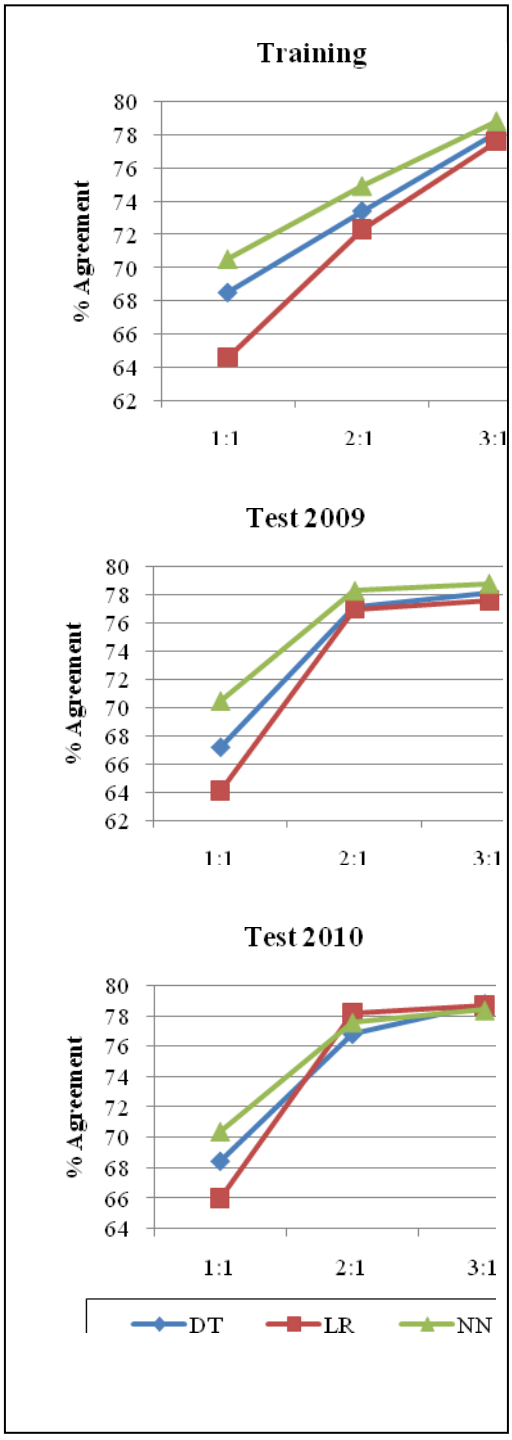


Fig. 1: The overall prediction accuracy rates of the prediction techniques over the three ratio datasets

We can see how much the difference is, and this holds for all of the prediction methods within all of the dependent variable values ratios and in the two years, 2009 and 2010, even for the training data. On the other hand, It is seen that for all of the prediction techniques and when the dependent variable values distribution was 1:1, the prediction accuracy rates for the "0" and "1" were highly comparable. The "0" and "1" values were fairly predicted by the prediction techniques with acceptable accuracy rate and the difference was very small compared with the difference when the ratio was 2:1 or 3:1. If we take the same example, the neural network in the year 2009 data and for the ratio 1:1, predicted the value of "0" and "1" for 70.8% and 69.6% respectively. This holds for all of the prediction methods within all of the dependent variable values ratios and in the two years, 2009 and 2010, even for the training data. The tradeoff in this situation was that the overall prediction accuracy rate was reduced significantly, but still acceptable, when the ratio became 1:1.

TABLE 5: PREDICTION ACCURACY PERCENTAGE OF THE PREDICTION TECHNIQUES FOR THE TESTING DATA 2009 AND 2010

Method*	Ratio	Test 2009			Test 2010		
		0:Employed	1:Unemployed	Overall	0:Employed	1:Unemployed	Overall
DT	**1:1**	**66.2**	**70.3**	**67.2**	**71.9**	**56.8**	**68.4**
DT	2:1	89.7	38.1	77.1	90.7	31.5	76.8
DT	3:1	95.8	23.4	78.1	96.9	19.7	78.8

LR	**1:1**	**63.7**	**65.3**	**64.1**	**67**	**62.6**	**66**
LR	2:1	93.1	27.2	77	93.8	27.4	78.2
LR	3:1	96.5	19.1	77.6	97.2	18.6	78.7
NN	**1:1**	**70.8**	**69.6**	**70.5**	**71.2**	**69.6**	**70.4**
NN	2:1	90.1	41.7	78.3	90.5	35.5	77.6
NN	3:1	94.9	28.9	78.8	94.4	26.2	78.4

*: DT: Decision Tree, LR: Logistic Regression, NN: Neural Network

5. Conclusion

In this paper we worked on achieving the objectives that can be illustrated by performing the prediction techniques performance comparisons of: Logistic Regression, Neural Network and Decision Tree using dataset of higher level of accuracy regarding the content. We tried to identify more precise predictors that significantly define and affect the output by using the correlation analysis but the results have demonstrated a very small differences, even it was more accurate without the correlation results. We think this is due to the small number of predictors, that limited our chances to get a highly correlated and robust training dataset.

As a conclusion for this research, the neural network achieved an overall prediction accuracy rate higher than the decision tree and logistic regression when the dependent variable values distribution ratio was 1:1. The prediction performance of the three prediction methods was almost the same and too close to each other when the dependent variable values distribution ratio were 2:1 and 3:1.

Another interesting conclusion is that a tradeoff should be performed, that whether the needed prediction accuracy is a high overall prediction accuracy rate; an adequate and comparable both "0" and "1" dependent value prediction accuracy rate; or a high single "0" or "1" dependent value prediction accuracy rate. If a high overall prediction accuracy is needed then the dependent variable values distribution in the training data should be skewed and the ratio of 0:1 occurrences (or 1:0) for the dependent variable values should be at least 2:1 or larger. This hold also if the requested high prediction accuracy is one of the two dependent variable values, not both, then it's distribution in the training data should be at least 2 occurrences or more against to 1 occurrence for the other value. An example of this case is the breast cancer diagnosing in women that a high prediction accuracy is needed to check if the patient is infected like the studies of Delen D. [2] and Bellaachia A. [3].

If both dependent variable values are requested to be predicted in comparable prediction accuracy rate, then the training data should not be skewed and the ratio of dependent variable values occurrences should be equal and no more than 1:1. This holds for all of the three prediction techniques and not affected by the total population size of the data (training or testing), because we tested this on another different datasets with total population sizes ranges between 3000 and 4000 instances. This is true if we consider our dataset, between 18,000 and 38,000 instances, as a large dataset but not necessarily a huge dataset.

This conclusion agrees, in general, with the results and conclusion of Lahiri

R. [1], but contradict with the failure of neural network to predict one of the dependent variable values. In this study it is seen that the neural network succeeded and even slightly outperformed the logistic regression and decision tree techniques in predicting the values of the dependent variable values, "0" and "1".

Finally, as a future work to increase the prediction accuracy, we would like to find other techniques that help in finding optimal choice of predictors and do the same study. Also the total population size is another future area of research to test if the huge dataset, with hundreds of thousands or even millions instances, affected the prediction accuracy of the aforementioned prediction techniques in this study.

References

[1] Lahiri R., *Comparison of Data Mining and Statistical Techniques for Classification Model,* A Thesis submitted to the graduate faculty of the Louisiana State University in partial fulfilment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences. (December 2006).

[2] Delen D., Walker G., and Kadam A., *Predicting breast cancer survivability: a comparison of three data mining methods*, Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[3] Bellaachia A. and Guven E., *Predicting Breast Cancer Survivability Using Data Mining Techniques*, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006).

[4] M. Panda and M. R. Patra, *A comparative study of data mining algorithms for network intrusion detection*, proc. of ICETET, India, 2008.pp.504-507. IEEE Xplore.

[5] Amooee G., Minaei-Bidgoli B. and Bagheri-Dehnavi M., *A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.),* IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[6] Ho Yu C., DiGangi S., Jannasch-Pennell A., and Kaprolet C., *A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year*, Journal of Data Science 8(2010), 307-325.

[7] Kumari M. and Godara S., *Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction*, International Journal of Computer Science and Technology (IJCST) Vol. 2, Issue 2, June 2011.

[8] Shailesh K R et. al., *Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay*, JPBMS, 2011, 7 (15).

[9] Ibrahim Z. and Rusli D., *Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression*, 21st Annual SAS Malaysia Forum, 5th September 2007.

[10] Nancy P. and Geetha Ramani R., *A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data*, International Journal of Computer Applications (0975–8887) Volume 32– No.8, October 2011.

[11] C. Deepa, K. Sathiya Kumari, and V. Pream Sudha, *A Tree Based Model for High Performance Concrete Mix Design*, International Journal of Engineering Science and Technology Vol. 2(9), 2010, 4640-4646.

[12] S. Shanthi and R. Geetha Ramani, *Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms*, International Journal of Computer Applications (0975–8887) Volume 35–No.12, December 2011.

Developing New Methods To Find The Number Of RAM Chips In The Memory Decoding To Construct The Required Memory Size

Lecturer Eng. Mohammad M Abu Omar
Al-Quds Open University
Palestine
momar@qou.edu

1. Abstract

Integrated –Circuit random access memory RAM chips are available in a variety of sizes. If the memory unit needed for an application is larger than the size of one chip, it is necessary to construct an array of RAM chips which includes a combined number of RAM chips. The problem here is how to determine the dimensions of the array of RAM chips. This paper develops new methods to find the number of RAM chips in the array of RAMS in order to obtain the required memory size.

2. Introduction

A memory unit is a collection of storage cells together with associated circuits needed to transfer information in and out of the device. Memory cells can be accessed for information transfer to or from any desired random location , so the name random-access memory, defined as RAM [1].

A memory unit stores the binary information in words, these words are groups of bits. Each word is an entity of bits that move in and out of storage as a unit.[5],[1].

The communication between a memory and its environment is achieved through data input and output lines, address selection lines, and control lines that specify the direction of transfer. All of these elements are shown in the following figure[5],[6]:

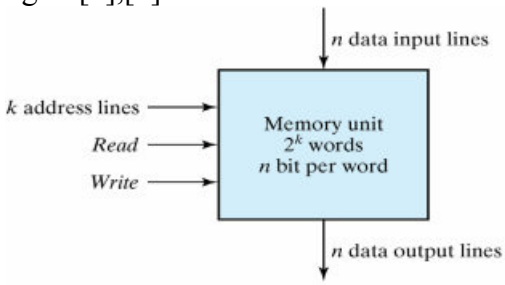


Figure 1 : Block Diagram of a memory unit

According to figure: 1, If there is a (2k * 4) RAM, this means:

(2k * 4) RAM = (2*2^10 * 4) = (2^11 * 4) RAM and from this we can conclude:

- The number of address lines = 11
- The number of words = 2^11
- The number of bits per word = 4

This research focuses on the communication between memory units in order to obtain a large memory size from small size memories .This is known and applicable by memory decoding winch includes an array of small size RAM chips to provide a large size of RAM., The new something that paper shows is how to determine the size of array of ram chips directly by knowing the size of small Ram and the wanted large size, It implements a mathematical equations that provide the following:

a- The number of rows in the array of Ram chips.
b- The number of columns in the array of Ram chips.
c- Also, the total number of Ram chips.
d- The type of the decoder used in the memory decoding.
e- The number of external logic OR – gates in the memory decoding.

3. Making Larger Memories

By using the CS lines, we can make larger memories from smaller ones by tying all address, data, and R/W lines in parallel, and using the decoded higher order address bits to control CS. Using the 4-Word by 1-Bit memory from before, we can construct a 16-Word by1-Bit memory. See the following figure [5],[6]:

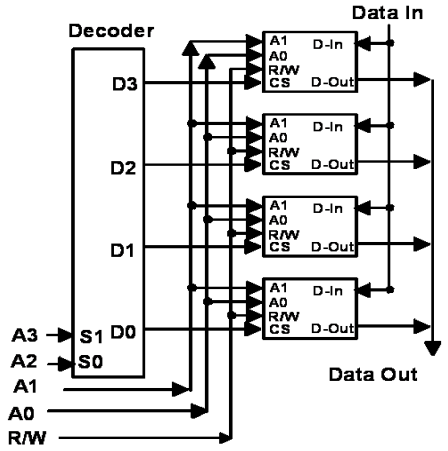


Figure 2 : Block Diagram of making a large memory

3.1 Making Wider Memories

To construct wider memories from narrow ones, we tie the address and control lines in parallel and keep the data lines separate. For example, to make a 4-word by 4-bit memory from 4, 4-word by 1-bit memories. Note that both 16x1 and 4x4 memories take 4-chips and hold 16 bits of data. See the following figure[5],[6]:

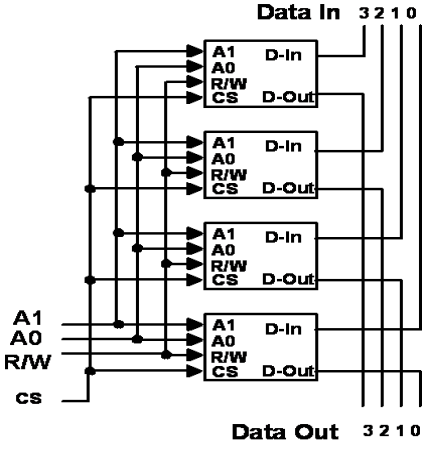


Figure 3 : Block Diagram of making a wider memory

3.2 The developed methods to find the number of RAM chips in the array of RAMS in order to obtain the required memory size

The methods of this research suppose the size of big Ram as :

X_1 U * N_1 RAM……..…(1)

And also suppose the size of small Ram as :

X_2 U * N_2 RAM…….…...(2)

Where:

X: is an integer number.

U: The unit of the size (ex: K: Kilo, M :Mega, G : Giga,..)

N: The number of bits in each word.

The number of rows in the array of RAM chips will be (R)

R = X_1 / X_2……..(3).

The number of columns in the array of RAM chips will be (C)

C = N_1 / N_2…..…….....(4)

The total number of small RAM needed will be calculated by

The following equation:

R * C …….....(5)

3.3 Using The Research Methods in Application Example

Suppose you need a memory array with 16k × 8 organization, but all you have on hand are 4k × 8 memory chips. How many 4k* 8 Ram chip needs ?

The size of big RAM is 16k * 16, By applying eqn (1) we obtain:

$X_1 = 16$ and $N_1 = 16$

The size of small RAM is 4k * 8, By applying eqn (2) we obtain

$X_2 = 4$ and $N_2 = 8$

Now :

The number of rows in the array of RAM chips

R = X_1 / X_2 , from eqn (3)

R = 16 / 4 = 4 Rows

The number of columns in the array of RAM chips

C = N_1 / N_2 , from eqn (4)

C = 16 / 8 = 2 Columns

The total number of small RAM chips needed

= R * C , from eqn (5)

= 4 * 2 = 8 chips from (4k * 8) RAM.

4K * 8 RAM	4K * 8 RAM
4K * 8 RAM	4K * 8 RAM
4K * 8 RAM	4K * 8 RAM
4K * 8 RAM	4K * 8 RAM

Figure 4 : Block Diagram of multiple 4K * 8 RAM

3.4 How to build a memory array from multiple Smaller RAMs to obtain Larger RAMs:

Here there is a necessary to find the suitable decoder, and also the number of external logic OR gates, the research methods in section 3.2 achieve these requirements.

From equation (3), R can help us to find the decoder type, Since the value of R here means the output of the decoder which equals 2 ^n , so we can determine the type of decoder which is (n * 2^n).

The number of OR gates can be obtained from equation (2), C

here refers to the number of OR gates needed.

3.4.a Application Example

Suppose you need a memory array with 8k × 16 organization, but all you have on hand are 2k × 4 memory chips. How many 2k* 4 Ram chip needs ? Show how you could connect them to form the desired array ?

The size of big RAM is 8k * 16, By applying eqn (1) we obtain:

$X_1 = 8$ and $N_1 = 16$

The size of small RAM is 2k * 4, By applying eqn (2) we obtain

$X_2 = 2$ and $N_2 = 4$

Now:

The number of rows in the array of RAM chips (R)

$R = X_1 / X_2$, from eqn (3)

$R = 8 / 2 = 4$ Rows

The number of columns in the array of RAM chips (C)

$C = N_1 / N_2$, from eqn (4)

$C = 16 / 4 = 4$ Columns

The total number of small RAM chips needed

= R * C , from eqn (5)

= 4 * 4 = 16 chips from (2k * 4) RAM

We can determine the decoder type from the value of R , eqn(3), the R equals 4, so the output of the decoder

2^n = 4 , and this implies that the input of the decoder is n = 2, this means that the decoder type is (2 * 4) decoder.

Also, we can determine the number of OR gates needed from the value of C eqn(4), so the number of OR gates = C = 4

OR gates.

From the previous analysis by using the research methods, we can draw the following block diagram of a 8k * 16 RAM :

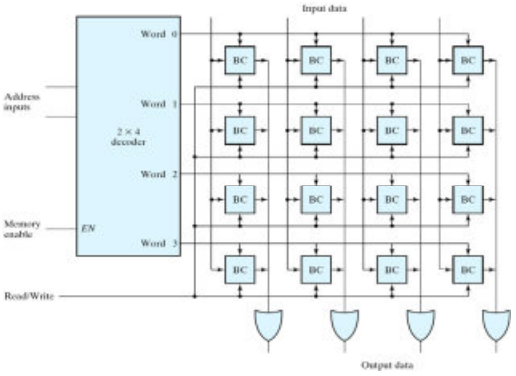


Figure 5: Block diagram of a 8k * 16 RAM

4.Conclusion

The constructing a big RAM size from multiple small RAM size is very important in memory organization, especially if there is only a small RAM size, and there is a need to a big size. The memory decoding solves this problem by constructing an array of RAM chips, which includes a combined number of small size RAMS to form the required memory size. This paper implements mathematical equations in order to determine the dimensions of the array of RAM chips, these equations provide the number of rows and columns in the array and also the total number of small RAM chips needed. On the other hand these equations may be used to determine the suitable decoder type and the number of OR logic gates which are used in the building of the array of small RAM chips.

References

[1] Mano,Morris,1990,Digital Design, Prentice-Hall

International Editions.

[2] Radhakrishanan, Rajaraman,2010,

An Introduction to Digital Computer

Design, Prentice –Hall International Editions.

[3] Radhakrishanan, Rajaraman,2006,

Digital Logic and Computer Organization,

Prentice –Hall International Editions.

[4] Radhakrishanan, Rajaraman,2007,

Computer Organization and Architecture,

Prentice –Hall International Editions.

[5] Mano Morris, Kime R Charles,2033,*Logic and Computer*

Design Fundamentals, Prentice-Hall International Editions.

[6] Kime R Charles, Kaminski Thomas,2008, *Logic and*

Computer Fundamentals.

http://writphotec.com/mano4/PowerPoint_Handouts/LCDF4_Chap_08.pdf.

[7] Singth, Avtar and Triebel,Walter A,1991,*16-Bit and 32-Bit*

Microprocessors, , Prentice-Hall International Editions.

[8] Groth, David,2033,*A+ Complete*, San Francisco. London.

[9] Memory and Programmable Logic. Lecture Notes

http://www.slidefinder.net/c/chapter_memory_programmable_logic/11304394

A social network algorithm for detecting communities from weighted graph in Web Usage Mining system

Yacine SLIMANI and Abdelouahab MOUSSAOUI
Laboratoire de Recherche en Informatique Appliquée LRIA – UFAS, Algeria
slimany09@gmail.com , moussaoui.abdel@gmail.com

Abstract

Web Usage Mining is the process of discovering user's navigation pattern and predicting user's behavior. The quantity of the Web usage data to be analyzed and its low quality are the principal problems in WUM. Several algorithms of data mining have been applied in order to extract the behaviors of the Web sites' users. In this present work, we have implemented a community detection technique in WUM process that is based on the modularity function and we have organized the preprocessed data as a weighted graph. The obtained results illustrate the aptitude of the proposed algorithm to determine a pertinent design of the web site from the discovered communities.

Keywords: *Data Mining, Web Usage Mining, log files, community discovery, weighted graph, social network, modularity.*

1. Introduction

One of the most significant axes of the Web Mining is the Web Usage Mining (WUM) which is interested in the extraction of the access pattern to the Web from the used data. The principal interest of the Web Usage Mining is that it provides information on the way in which the users browse the Web site [1].

In this work, we are interested in the analysis of the user browsing behavior. The objective is to understand the navigational practices of users (teachers, students and administrative staff).

Cooley [2] divides the WUM in three main steps: preprocessing, pattern discovery and pattern analysis. The preprocessing task within the WUM process involves cleaning and structuring data to prepare it for the pattern discovery task. In the phases of discovered and analyzes knowledge, the Web Usage Mining represents a field of research to discover the behavioural models of the users [3].

In our work, we have first cleaned the data by removing no relevant information and the noise. The remaining data are arranged in a

coherent way in order to identify, in a precise way, the users sessions.

We then defined a new approach of extraction which treats the data resulting from the preprocessing phase as being a set of communities. Our aim is to extend the application of the recent community detection methods in the Web Mining context in order to profit from their classifying capacity in the communities discovery.

The rest of the paper is organized as follows. Section 2 describes the Web usage data preprocessing which we intend to increase the quality of the data obtained at the end of the preprocessing step. In section 3, we present an approach that extract interesting correlations from the data based on discovery community method. Section 4 contains our experimental results. General remarks and conclusions are presented in section 5.

2. Preprocessing method

The generic process WUM is adapted to each axis of the Web mining according to the nature of the used data (text, logs, edges…). The functional structure of the process of the web usage mining is structured in six modules principal like representing in figure 1.

2.1 Data transformation module

The entry of the data transformation module is a log file which is a textual file that records the requests made to the Web server in chronological order. The most used formats for log files are CLF (Common Log Format) and the ECLF (Extended CLF). We use the standard ECLF. An example of this format is as follow:

41.200.89.109 - - [12/Oct/2008:20:18:23 +0100] "GET/citic2008/soumission.html HTTP/1.1" 200 23247

"http://www.univ-setif.dz/citic2008/index.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; fr; rv:1.9.0.3)

Gecko/2008092417 Firefox/3.0.3

1) the name or IP address of the appealing machine.
2) the name and the login HTTP of the user.
3) the date and the hour of the request.
4) method used by the request (Get, Post, etc.)
5) the URL of the request.
6) the used Protocol.
7) the request statute .
8) size of the sent file.
9) the URL which referred the request.
10) the Agent (navigator and the operating system)

The analysis of Web log files permits to identify useful patterns of the browsing behavior of users which can be exploited in the process of Web personalization.

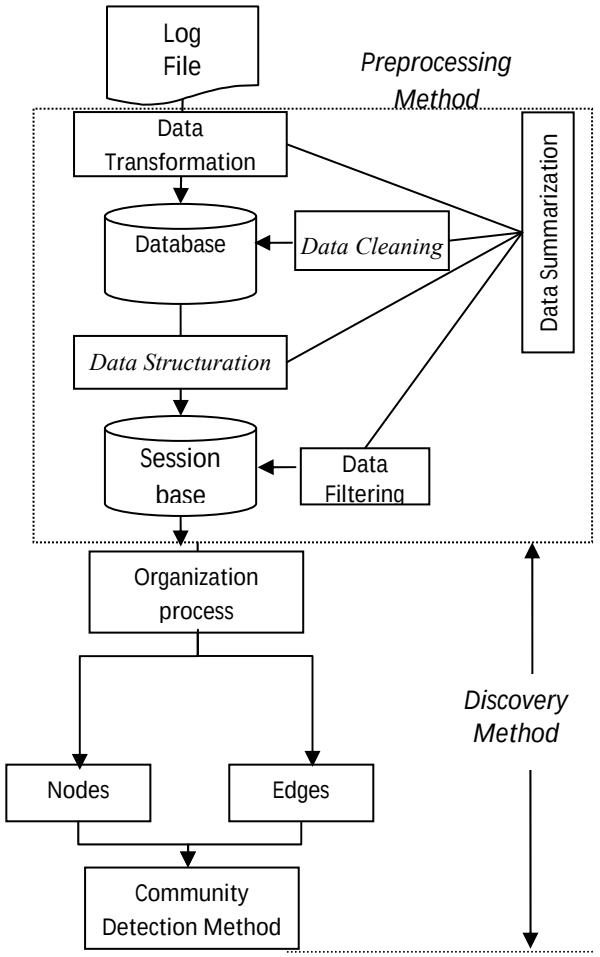


Fig.1 Architecture of the web Usage Mining Process.

2.2 Data cleaning module

The data cleaning module is used to remove the useless records in order to maintain only users' data which can be accurately exploited to identify browsing behavior of users. The choice of the data to be removed depends on the ultimate objective of the personalization system of the Site. In our work, the objective is to develop a WUM system to offer personalized dynamic links to the site's visitors. Ttherefore the system has to keep only records relating to explicit requests that represent users' actions. Consequently, the data cleaning module was developed to eliminate the following requests:

2.2.1 Method different from "GET"

In general, the requests containing a value different from "GET" are not explicit requests of the users, but they often relate to accesses with CGI, of the visits of robots, etc. Consequently, these requests are regarded as non significant and are withdrawn from the access log files.

2.2.2 Failed and corrupted requests

These requests are represented by records containing a HTTP error code. A status with value different from 200 represents a failed request (e.g. a status of 404 indicates that the requested file was not found at the expected location).

2.2.3 Requests for multimedia objects

In the HTTP protocol, an access request is carried out for every file, image, multimedia object embedded in a requested Web page. As a consequence, a single request for a Web page may often produces several entries in the log file that corresponds to files automatically downloaded without an explicit request of the same user. The requests of this type of files can be easily identified since they contain a particular URL name suffix, such as gif, jpeg, jpg, and so on. The conservation or removal of these multimedia objects depends on the kind of the Web site to personalize and theirs natures. In general, these requests do not represent the effective browser activity of the user visiting the site, hence they are removed. In other cases, eliminating requests for multimedia objects may cause a loss of useful information.

2.2.4 Requests originated by Web robots

Log files contain some number of records corresponding to requests originated by Web robots. Web robots are programs that automatically download complete Web sites by following every hyperlink on every page within the site in order to update the index of search engine. These requests are not regarded as usage data and, consequently, have to be removed. To identify web robots' requests, the data cleaning module implements two different heuristic [4].

Firstly, all records containing the name "robots.txt" in the requested resource name (URL) are identified and removed. The second heuristic is based on the fact that web robots retrieve pages in an automatic and exhaustive manner, so they are characterized by a very high browsing speed that is equal to total number of visited pages / total time spent to visit those pages. Therefore, for each different IP address we calculate the browsing speed and all requests having this value exceeding a threshold (pages/second) are regarded as made by robots and are consequently removed. The threshold value is determined after reviewing the log files. After data cleaning, only requests for relevant resources are saved in the database. At the end of this step, we formally define $R = \{r_1, r_2, ..., r_{n_r}\}$ as the set of all distinct resources requested from the Web site under analysis.

2.3 Data structuration module

The data structuration module regroups requests of the log file in user sessions. A session is defined as a limited set of resources accessible by the same user within a particular visit. The identification of user sessions from the log data is a difficult task because many users can use the same computer and the same user can use different computers. Therefore, one main problem is how to identify the user. For websites that require user registration, the log file contains the user login that can be used for user identification. When the user login is not available, the user is identified from the IP address, i.e. we consider each IP address as a different user (being aware that an IP address might be used by several users) [5].

We define $U = \{u_1, u_2, ..., u_{n_u}\}$ as the set of all the users that have accessed that website. We use a time-based method to identify sessions [2] [8]. A user session represents the set of all access originating from the same user within a predetermined time. This period is determined by considering a maximum elapsed time $\Delta t_{\max}$ between two consecutive accesses. Moreover, to better handle special situations for example, when users access several times to the same page due to the slow connections or intense network traffic, a minimum elapsed time $\Delta t_{\min}$ between consecutive accesses is also fixed [4]. We define a user session as:

$s^{(i)} = (u^{(i)}, t^{(i)}, r^{(i)})$ Where:

$u^{(i)} \in U$: is the user identification.

$t^{(i)}$: is the access time of the whole session.

$r^{(i)}$: is the set of all resources requested during the i-th session (with corresponding access time), namely:

$$r^{(i)} = ((t_1^i, r_1^i), (t_2^i, r_2^i), ..., (t_{n_i}^i, r_{n_i}^i)) \qquad (1)$$

with $r_j^i \in R$

Where access time t_k^i to a single resource satisfies the following:

$t_{k+1}^i \geq t_k^i$ and

$\Delta t_{\min} < t_{k+1}^i - t_k^i < \Delta t_{\max}$

Summarizing, after the data structuration phase, a set of n sessions $s^{(i)}$ is identified from the log data. We denote the set of all identified sessions by: $S = (s^{(1)}, s^{(2)}, ..., s^{(n_s)})$.

Once all sessions have been identified, the data structuration module presents a panel that lists the extracted sessions and allows us to view and save the details (IP address, requested resources in the session, date and time of the requests) of each user session.

2.4 Data filtering module

After the identification of user sessions, we perform a data filtering step to remove the less requested resources and retain only the most requested ones. For each resource r_i, we consider the number of sessions NS_i that required the resource r_i, and we compute the quantity $NS = \max_{i..n_R} NS_i$. Then, we define a threshold ε, and we remove all request with $NS_i < \varepsilon$ are removed. In this way, the data filtering module can significantly reduce the number of relevant requested resources, which facilitates treatment of the next phases of the web usage mining.

2.5 Data Summarization Module

The Data Summarization Module generates reports summarizing the information obtained after the application of pre-processing step. This statistical information permit to obtain a schematic and concise description of the usage data mined from the analyzed log file. It provides the necessary information to detect some particular aspects related to the user browsing behavior or to the traffic of the considered site log file.

3. Discovery method

Once the raw logs have been preprocessed, data mining techniques can be applied on the dataset to discover new patterns. Such techniques include, but are not limited to: association rules mining, sequential pattern mining and clustering. In our work, we have suggested the use of the recent method of community detection in order to identify groups of users with similar behavior for which personalized versions of the Web site may be created.

In the second phase of WUM process and in order to find a pattern discovery, we have applied an organization process which consists in analyzing the pretreated data of the session base and to model them via a functional graph, such as the resources will be represented by nodes and the browsing sequences of users during each session will be represented by edges. After obtaining this graph, we proceed to the identification of the users clusters which have similar behaviors in term

of visited content, our choice is based on Newman algorithm [6] and the modularity function [9] to identify the community structure and thus to define the suitable pattern discovery.

3.1 Concepts of community structure

In complex networks, the communities are groups of nodes which share probably a common proprieties and/or similar functions. The communities may be correspond , for example, to groups of Web pages accessible over the Internet that have the same subject [10], functional modules as cycles and pathways in metabolic networks[11], a set of people or groups of people with some pattern of contacts or interactions between them [12, 13], and subdivisions in the food webs [14,15]. In this paper, the communities correspond to groups of web pages which show the same browsing behavior of users. Newman and Girvan [16] introduce a measure of quality of a particular partition which they called “modularity” to detected if communities are good or no and to value such partitions. The modularity is based on assortative mixing measure [17].

Modularity measures when the division is a good one, in the sense that there are many edges within communities and only a few between them.

3.2 Organization process

Once user sessions have been identified, we have to use them to extract the Web graph that represents the analytic network. We have applied an algorithm that can be used to obtain the degree and the edges for a Web resource. In WUM process, we need to treat the case of weighted graph because resources may be visited several times in the same session and also through the different sessions. In fact, as can be seen in figure 2, we have computed the degree of resource as strictly related to the frequency of accesses to that resource and we have determined the whole weighted social network.

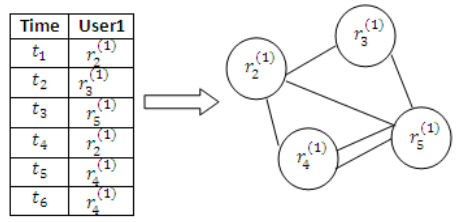


Fig. 2 *Organization process*

In this example the result is a weighted network containing far more information than a simple binary adjacency matrix. Thus the

adjacency matrix represents the weight of connection from r_v to r_w .

We could obtain insight into the behavior of weighted graphs very simply by mapping them onto unweighted multigraphs and any techniques that can normally be applied to unweighted graphs can be applied to the multigraph as well [19].

3.3 Pattern discovery task

Naturally, in addition to dividing the graph top down into clusters, one may also work bottom up merging singleton sets of nodes iteratively into clusters. Such methods are called agglomerative clustering algorithms. In our method, we have used the fast algorithm [6] and we have applied the idea presented in [19] in order to focus on weighted social network.

Let $G=(V,E)$ be a weighted graph describing a pretreated database of session with V the set of nodes and E the set of edges. We define the Modularity function as it is define in [18], thus the is

$$Q=\frac{1}{2m}\sum_{r_v r_w}\left[A_{r_v r_w}-\frac{k_{r_v}k_{r_w}}{2m}\right]\delta(c_{r_v},c_{r_w}) \qquad (2)$$

Where m denotes the total number of edges of the graph, so $m=\frac{1}{2}\sum_{r_v r_w} A_{r_v r_w}$

k_{r_v} is the degree of node r_v, we write $k_{r_v}=\sum_{r_v} A_{r_v r_w}$

The element $A_{r_v r_w}$ of the adjacency matrix of the network represents the weight of connection from r_v to r_w. The nodes are dived into communities such that node r_v belongs to community c_{r_v} and the $\delta-$function yields one if nodes r_v and r_w are in the same community, zero otherwise.

The steps of the algorithm are as follows.

1. We construct the weighted adjacency matrix $A_{r_v r_w}$ of the analyzed graph *G*.
2. Initially, we consider that each community is composed of a single node.
3. Joining the pair of communities whose amalgamation produces the largest increase in $\Delta Q_{r_v r_w}$ but do not join the pair of communities whose there are no edges between them. The $\Delta Q_{r_v r_w}$ value is written as follows

$$\Delta Q_{r_v r_w} = \begin{cases} \frac{1}{2m} - \frac{k_{r_v} k_{r_w}}{(2m)^2} & \text{if } r_v, r_w \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

4. We update the elements of the matrix $\Delta Q_{r_v r_w}$ and the modularity matrix $Q_{r_v r_w}$ in the manner that they correspond to the joined communities.

5. Repeat step 3 until only one community remains.

For a network of n vertices, after $(n-1)$ such joins we are left with a single community and the algorithm stops. The partition corresponding to the maximum value of modularity on this graph should be the best or at least a very good one.

4. Analysis and experiments

4.1 Preprocessing results

Our preprocessing method has been tested on log files stored by the Web site Server of Ferhat Abbas University of Setif (Algeria) available at the URL www.univ-setif.dz. The treated file covers the site activities during the period from 17 January 2010 to 14 February 2010. Table I presents the results of the preliminary analysis of log files and synthesizes the results provided during the data summarization phase.

TABLE V
THE DATA TRANSFORMATION SUMMARY.

File size	100 448 034 bytes
Date/ beginning hour	17/01/2010 04:03:30
Date/ending hour	14/02/2010 09:09:00
Number of requests	365 863

In the following, we analyze the log file, during the data cleaning phase, in order to determine the non explicit user requests.

The number of requests corresponding to multimedia objects is very important. They include 72,48 % of the requests. After data cleaning phase, only 27,52% of the requests are maintained in the database.

TABLE VII
THE DATA CLEANING SUMMARY.

Multimedia		**Method**		**Status**	
Autre	100 691	Get	363 749	200	257 276
.gif	53 177	!get	24	206	20 038
.png	88 413	head	952	301	512
.jpg	69 449	options	290	304	75 776
.ico	12 005	post	705	400	22
.bmp	176	propfind	93	403	226
.css	17 866	put	50	404	11 927
.js	24 077			405	62
				501	24
Total	**265 172**	**Total**	**2 114**	**Total**	**108 587**
72,48%		**0,58%**		**29,68%**	

In case of Get method, we remove the requests that have access methods different from it. Table 2 presents the number of each type of method in the log file. We note that the number of removed requests is very small

compared to the total number (0,58%).

The requests which have a status different from 200 are regarded as failed request. We list three major categories of irrelevant requests: 3% of the requests with a status of 404 indicate that the requested file was not found at the expected location, 5% of the requests with a status of 206 and 20% of the requests with a status code of 304 indicate that the requested file have a browser refresh problem. At the end of the data cleaning phase, we retain 70% of the requests.

Table II recapitulates the different removed requests of the database. We observe that overlapping can occur between two removed categories. For example, a request with method "Head" can also be a request for a multimedia object. In this case, the data summarization module counts twice the removal of the request, even though only one record is deleted from the log file. Table III illustrates this overlapping.

TABLE VIIII
OVERLAPPING BETWEEN THE REMOVED REQUESTS CATEGORIES.

Request category				Nb	%
	Multi media	Method ≠ Get	Status ≠ 200		
Cleaned	X			183 583	50,18
		X		1 579	0,43
			X	26 548	7,26
	X	X		25	0,01
	X		X	81 529	22,28
		X	X	475	0,13
	X	X	X	35	0,01
Valid				72 089	19,70
Total	265 172	108 587	2 114	365 863	100 %
%	72,48%	29,68%	0,58%	100 %	

After the data cleaning phase, the log files were processed by the data structuration module in order to define the two sets R and U :

R The whole of the requested resources from the analysed web site.

U The web site users.

Then we apply the structuration algorithm [7] to determine the sessions taking account of the two values: $\Delta t_{\max}$ and $\Delta t_{\min}$, such as the minimal value is used to detect the robots and the web crawler, and the maximal value allow detection of new sessions (Table IV).

TABLE VIIIV

THE STRUCTURATION DATA SUMMARY

Input data	
number of requests	103975
Nb IP (U)	8 676
Nb URL (R)	7 707

Setting	
$\Delta t_{\max}$	30 minutes
$\Delta t_{\min}$	05 seconds
Nb Sessions:	17 379

The last phase of the preprocessing method is resources filtering, we have removed the least requested URLs according to the ε threshold and we have obtained the results illustrated in Table V.

TABLE V

THE FILTERING DATA SUMMARY.

	Before	ε = 5	ε = 50	ε = 100
Nb Req.	103 975	66 792	48 571	29 481
Nb URL	7 707	1 607	193	105
Nb IP	8 676	2 843	2 829	1 480
Nb Sess.	17 379	4 665	4 571	2 199

4.2 Community detection results

In the last phase of our work, we have applied an organization process to extract the functionally graph from the session base. So, we have obtained the network structure that identifies the users' session and all the sessions (the nodes represent resources and edges represent the browsing sequences of users during each session). Any community detecting algorithm requires establishing its analytical network. Figure 3 shows this network structure.

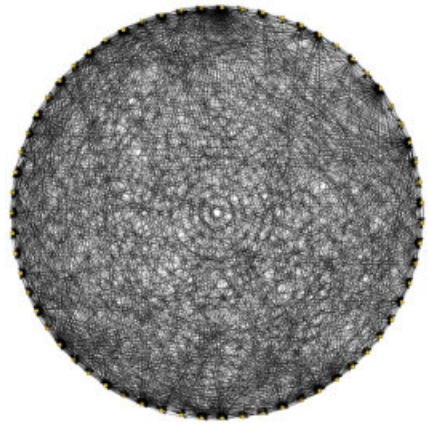


Fig. 3 Network structure.

In pattern discovery step, we intend to identify community structure and detect the browsing behavior of users which can be exploited in the process of Web personalization. A community structure is a set of nodes which have more internal density within the community than with the rest of the network [17]. The proposed discovery method belongs to hierarchical partitioning approach to clustering. It produces a nested sequence of partitions of the set of data points, the used web graph contain 66 nodes and 1180 edges (table 6), which can be displayed as a tree with a single cluster, including all points at the root and singleton clusters (individual points) at the leaves.

TABLE VI

DATA INPUT OF THE USED WEB GRAPH.

Number of nodes	Number of edges
66	1180

The detection community algorithm computes $\Delta Q_{r_v r_w}$ and find the pair of communities r_v, r_w with the largest $\Delta Q_{r_v r_w}$. The output of the algorithm can be represented in the form of a dendrogram (Fig. 4) and the optimal section of the dendrogram found by looking for the optimal value of Q (Fig. 5).

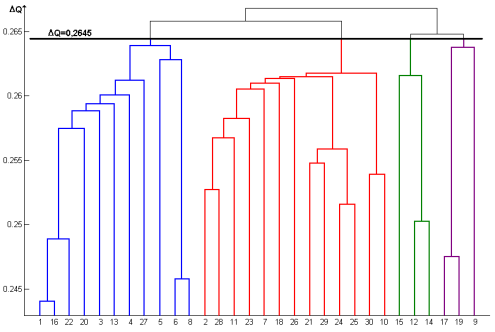


Fig. 4 The dendrogram represent the partitioning of network

As is known to all, modularity Q with the maximum value corresponds to the best partition of community detecting, here the best value that we have obtained is 0.2645.

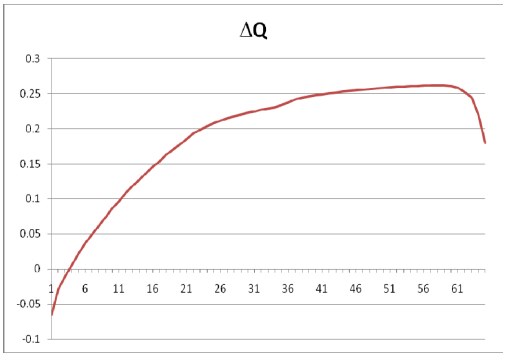


Figure 4. Value of modularity.

By applying the pattern discovery method, we have obtained 4clusters. The cluster number 3 contains the visits to the Web pages of scientific event (e.g. call of paper, program and important dates). In this case, the goal of these users is precise: the visitor interest is to consult the scientific activities of Ferhat Abbas University, the cluster number 4 regroups the visits interested in Web pages of research and valorization, the first cluster detects the visits between the deferent galleries of images and the second cluster shows the visits to deferent faculties. These analyses have permitted to identify homogenous classes of visitors. The graph obtained is presented in Fig. 6.

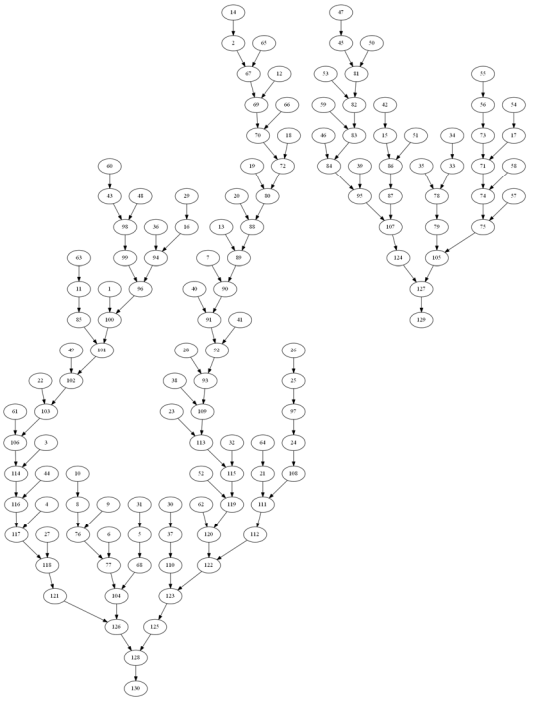


Figure 6. The 4 classes of visitors

5. Conclusions

The Web Usage Mining can be improved by using, in the two steps of the WUM process, enriched information about the structure and content of the Web sites analyzed. The preprocessing method that we have used allows a significant reduction in the number of initial requests and offers a structured session base for the next step of discovery method. The implemented discovery algorithm takes into account more suitable information given by the weighted graph.

References

[1] D. Tanasa. Web usage mining : Contributions to intersites logs preprocessing and sequential pattern extraction with low support. Ph. D. Thesis, University of Nice Sophia Antipolis, 2005.

[2] R. Cooley. Web usage mining : Discovery and application of interesting patterns from web data. Phd thesis, University of Minnesota, 2000.

[3] Pierrakos D. et al. Web usage mining as a tool for personalization: a survey. User Modeling and User-Adapted Interaction, 13(4), p. 311-372 (2003).

[4] Tan,P. N. and Kumar, V.. Discovery of Web Robot Sessions Based on their Navigational Patterns, Data Mining and Knowledge Discovery, 6(1), p. 9 -35 (2002).

[5] Suryavanshi B.S. et al. A Fuzzy Hybrid Collaborative Filtering Technique for Web Personalization. In Proc. of 3rd Workshop on Intelligent Techniques for Web Personalisation (ITWP'05).

[6] Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Physical Review E, Vol 70, 066111

[7] Nasraoui O. World Wide Web Personalization. In J. Wang (ed), Encyclopedia of Data Mining and Data Warehousing, Idea Group (2005).

[8] Paliouras G. et al. Large-scale mining of usage data on Web sites. AAAI Spring Symposium on Adaptive User Interface, Stanford, California, p.92-97 (2000).

[9] Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Physical Review E, Vol 70, 066111

[10] Flake GW, Lawrence S, Lee Giles C, Coetzee FM, Self-Organization and Identification of Web Communities. IEEE Computer, Vol 35, No 3, pp 66–71,2002.

[11] Guimer R, Amaral LAN, Functional cartography of complex metabolic

networks. Nature 433, pp 895-900, 2005.
[12] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. PNAS, 99(12) :7821-7826, 2002.
[13] Lusseau D, Newman MEJ (2004), Identifying the role that animals play in their social networks. Proceedings of the Royal Society of London B, Vol 271, pp S477–S481.
[14] Pimm SL (1979) The structure of food webs. Theoretical Population Biolology,Vol 16, pp 144–158.
[15] Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) , Compartments exposed in food-web structure. Nature, Vol 426, p 282–285.
[16] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113,2004.
[17] M. E. J. Newman, Mixing patterns in networks, Phys. Rev. E 67, 026126 - 2003.
[18] M. E. J. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E* 69, 066133-2004.
[19] M. E. J. Newman, Analysis of weighted networks, *Phys. Rev. E* 70, 056131 (2004).

Applying Data Mining Technology in Modeling and Predicting Number of Students in Bedia Center

Ola Rayyan
AL-Quds Open University, Palestinian
orayyan@qou.edu

Abstract

In this document we review the concept of Data Mining, and we will show how this technology help in taking decisions base on historical stored data. Data Mining uses several kinds of modeling techniques, and we will focus on Regression Model technique which is used to predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric. The forms of regression are linear, multiple, weighted, polynomial, nonparametric and robust.

In this case study we will predict and estimate the number of students will enroll in Bedia Educational Center which is branch of Al-Quds Open University. The estimation is based on the main resource of data which is the historical data for those students were enrolled in Salfeet Educational Region; it is also another branch for Al-Quds Open University. This prediction and estimation based on simple linear regression modeling technique. The reason of using this technique is the steady increasing in the number of students at the university in general and at Salfeet Educational Region in specific. This case study will give the decision makers at Al-Quds Open University view about the number of students in the future, which help them to take the right decision for the situation of Bedia Educational Center. Also they can manage its arrangements with more precise estimations such as revenues, expenses and employment stuff.

***Keywords*:** *Data Mining, Regression, Simple Linear Regression*

1. Introduction

AL-Quds Open University is distributed university through the Palestine, because it is university for open education. This type of universities required to be spread at all districts of Palestine. The requirements to open new educational branch depends on the number of students will enroll in this new branch. For this reason we need to

estimate the number of students based on the previous historical data for another educational region like Salfeet Educational Region using regression model.

The main problems is learning the regression model, spend more time to learn how to use the modeling tool like SPSS. Also exporting the data from the database and transforming it to suitable my work and aggregate it to do the mining analysis for the data.

Regression models are used to predict one variable from one or more other variables. Regression models provide the scientist with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events. The scientist employs these models either because it is less expensive in terms of time and/or money to collect the information to make the predictions than to collect the information about the event itself, or, more likely, because the event to be predicted will occur in some future time.

In order to construct a regression model, both the information which is going to be used to make the prediction and the information which is to be predicted must be obtained from a sample of objects or individuals. The relationship between the two pieces of information is then modeled with a linear transformation. Then in the future, only the first information is necessary, and the regression model is used to transform this information into the predicted. In other words, it is necessary to have information on both variables before the model can be constructed.

For example, the personnel officer of the widget manufacturing company might give all applicants a test and predict the number of widgets made per hour on the basis of the test score. In order to create a regression model, the personnel officer would first have to give the test to a sample of applicants and hire all of them. Later, when the number of widgets made per hour had stabilized, the personnel officer could create a prediction model to predict the widget production of future applicants. All future applicants would be given the test and hiring decisions would be based on test performance.

A notational scheme is now necessary to describe the procedure:

Xi is the variable used to predict, and is sometimes called the independent variable. In the case of the widget manufacturing example, it would be the test score.

Yi is the observed value of the predicted variable, and is sometimes called the dependent variable. In the example, it would be the number of widgets produced per hour by that individual.

Y'i is the predicted value of the dependent variable. In the example it would be the predicted number of widgets per hour by that individual.

The goal in the regression procedure is to create a model where the predicted and observed values of the variable to be predicted are as similar as possible. For example, in the widget manufacturing situation, it is desired that the predicted number of widgets made per hour be as similar to observed values as possible. The more similar these two values, the better the model. The next section presents a method of measuring the similarity of the predicted and observed values of the predicted variable.

In This report we will talk about the background of data mining, and we will also talk about the case study and tools used in our work. We will discuss the result and evaluate these results.

2. Background

2.1 What Motivated Data Mining? Why Is It Important?

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining)[1].

2.2 What Is Data Mining?

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both "data" and "mining" became a popular choice. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging[1]. Many people treat data mining as a synonym for another popularly used term, \Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process is depicted in Figure 1.2, and consists of an iterative sequence of the following steps:

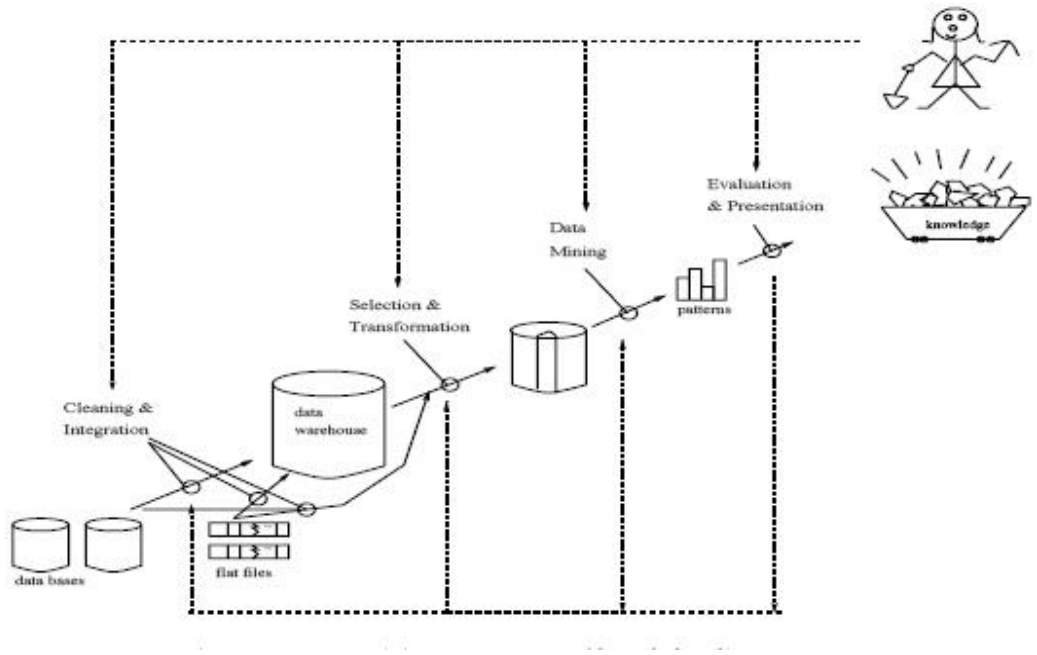


Figure 1.2: Data mining as a process of knowledge discovery.

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources maybe combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some **interestingness measures.**

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

2.3 Data Mining-On What Kind of Data?

In this section, we examine a number of different data stores on which mining can be performed. In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World Wide Web. Advanced database systems include object-oriented and object-relational databases, and specific application- oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

2.3.1 Relational Databases

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. The software programs involve mechanisms for the definition of database structures; for data storage; for concurrent, shared, or

distributed data access; and for ensuring the consistency and security of the information stored, despite system crashes or attempts at unauthorized access[2].

2.3.2 Data Warehouses

Data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data transformation, data integration, data loading, and periodic data refreshing.

2.3.3 Transactional Databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (Trans ID), and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person and of the branch at which the sale occurred, and so on.

2.3.4 Advanced Database Systems and Advanced Database Applications

Relational database systems have been widely used in business applications. With the advances of database technology, various kinds of advanced database systems have emerged and are undergoing development to address the requirements of new database applications.

The new database applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), time-related data (such as historical records or stock exchange data), and the World Wide Web (a huge, widely distributed information repository made available by the Internet). These applications require efficient data structures and scalable methods for handling complex object structures, variable-length records,

semi structured or unstructured data, text and multimedia data, and database schemas with complex structures and dynamic changes[3].

2.4 Data Mining Functionalities-What Kinds of Patterns Can Be Mined?

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions[4].

2.4.1 Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. These descriptions can be derived via (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms, or (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

2.4.2 Association Analysis

"What is association analysis?" Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

2.4.3 Classification and Prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

"How is the derived model presented?" The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.

2.4.4 Cluster Analysis

"What is cluster analysis?" Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the interclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

2.4.5 Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud

detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

2.4.6 Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

3. Case Study

3.1 Prediction number of student's case study

In this case study we will try to find a model to estimate the number of students for Bedia Educational Center based on the number of students on at Salfeet Educational Region those from Bedia region. We will use the simple linear regression in this case study.

3.2 Linear Regression and Prediction

Linear regression uses the relationship between distributions of scores in making predictions. If there is a relationship between two distributions, it is possible to predict a person's score in one distribution on the basis of their score in the other distribution (e.g., using a score on an aptitude test to predict actual job performance). Simple regression refers to the situation where there are only two distributions of scores, X and Y. By convention, X is the predictor variable, and Y the criterion (or predicted) variable.

3.2 Definitions

a) A **scatterplot** is a graph of paired X and Y values

b) A **linear relationship** is one in which the relationship between X and Y can best be represented by a straight line.

c) A **curvilinear relationship** is one in which the relationship between X and Y can best be represented by a curved line.

d) A **perfect relationship** exists when all of the points in the scatter plot fall exactly on the line (or curve). An **imperfect relationship** is one in which there is a relationship, but not all points fall on the line (or curve).

e) A **positive relationship** exists when Y increases as X increases (i.e., when the slope is positive).

f) A **negative relationship** exists when Y decreases as X increases (i.e., when the slope is negative).

3.3 Equation for a straight line

The equation for a straight line is usually written as:

$$Y=bX+a$$

where b = slope of the line

$= (Y2 - Y1) / (X2 - X1)$

= "the rise" divided by "the run"

and

a = the Y-intercept

= the value of Y when $X = 0$

Perhaps an example will help to clarify what this means. Imagine that you have decided to start working out at a gym. The annual membership fee is £25, and in addition to that, you must pay £2 every time you go to the gym.1 If we let X = the number of times you go to the gym, and Y = the total cost, we would find that:

$$Y=2X+25$$

The Y-intercept is 25. That is, if you never go to the gym (X = 0), your total cost is £25. And the slope of the line (b) is 2: Every time you go to the gym, it costs you another £2. Putting this another way, every time there is an increase of 1 on the X-axis, there is an increase of 2 on the Y axis.

3.4 Miscellaneous points about linear regression

Linear regression is used to predict a Y score from a score on X. Bear in mind the following:

1) The relationship between X and Y must be **linear**. If the relationship is not

linear, prediction will not be very accurate.

2) Normally, we are not interested in predicting Y scores that are already known. We derive our regression equation with sample data that consists of paired X and Y scores, but use the equation to predict Y scores when only X values are given. Because we use data collected from a sample to make these predictions, it is vital to have a **representative** sample when deriving a regression equation.

3) A regression equation is properly used only for the range of the variables on

which it was based. We do not know whether the relationship between X and Y

continues to be linear beyond the range of sample values.

4) Prediction is most accurate if the data have the property of **homoscedasticity**

i.e., if the variability of the Y scores is constant at all points along the regression

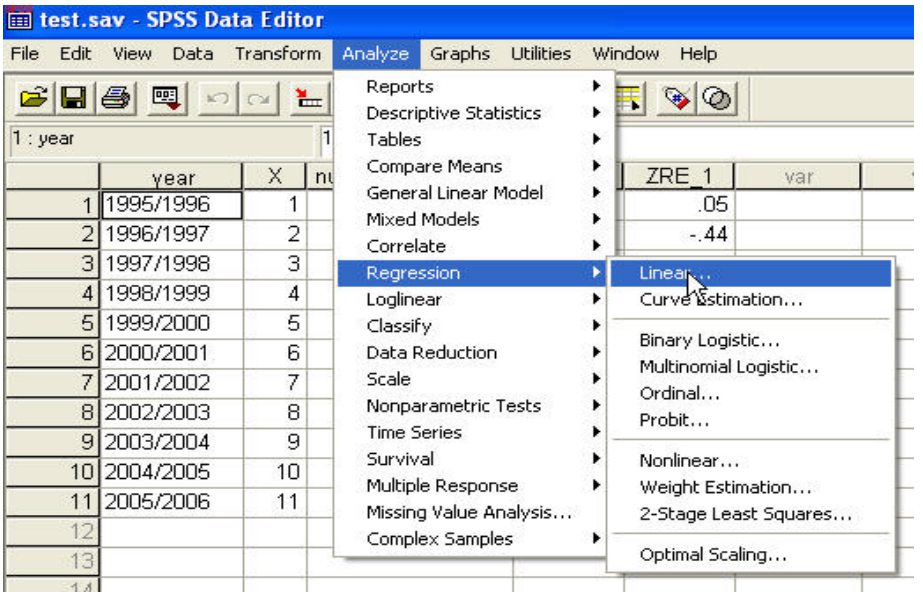
line.

5) When X and Y are both normally distributed and the number of paired scores is large, the data in a bivariate frequency distribution often produce a so-called

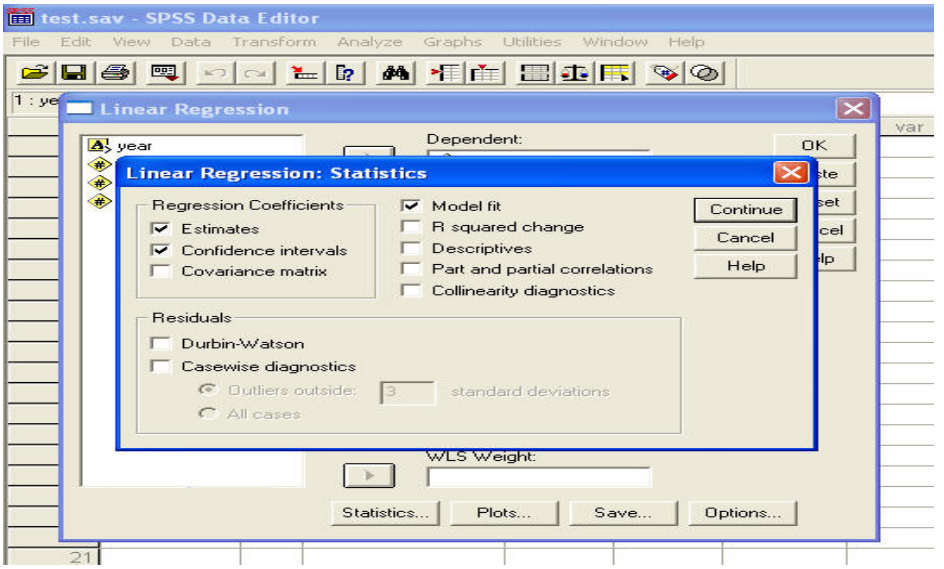
bivariate normal distribution. When you have such a distribution, the **standard error of estimate** can be used in the same way we used the standard deviation of a normal distribution. That is, we could say that about 68% of the scores in the scatterplot fall within 1 standard error of the regression line; and about 95% of the scores fall within 2 standard errors of the regression line.

3.5 Regression Analysis Using SPSS

The REGRESSION command is called in SPSS as follows:



Selecting the following options will command the program to do a simple linear regression and create two new variables in the data editor: one with the predicted values of Y and the other with the residuals.



The output from the preceding includes the correlation coefficient and standard error of estimate.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.969	.965	41.639

a. Predictors: (Constant), X

b. Dependent Variable: numberOfStudents

The regression coefficients are also given in the output.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	383.714	26.927		14.250	.000	322.801	444.627
	X	66.286	3.970	.984	16.696	.000	57.305	75.267

a. Dependent Variable: numberOfStudents

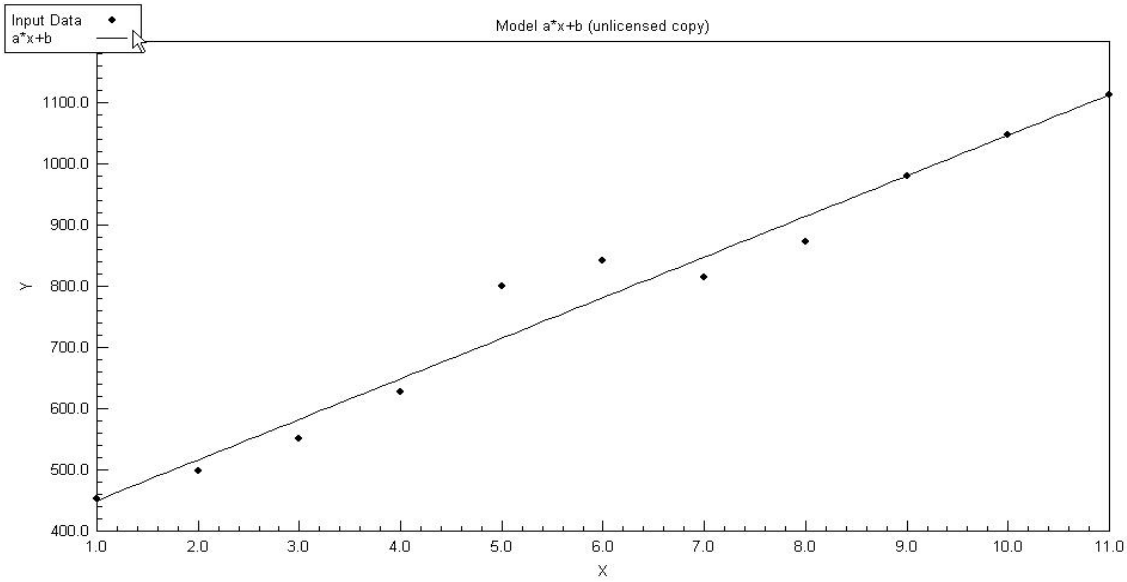
The optional save command generates two new variables in the data file.

ile Edit View Data Transform Analyze Graphs Utilities Window Help

: numberOfStudents 800

	year	X	numberOfStudents	PRE_1	ZRE_1	var
1	1995/1996	1	452	450.00	.05	
2	1996/1997	2	498	516.29	-.44	
3	1997/1998	3	550	582.57	-.78	
4	1998/1999	4	628	648.86	-.50	
5	1999/2000	5	800	715.14	2.04	
6	2000/2001	6	842	781.43	1.45	
7	2001/2002	7	814	847.71	-.81	
8	2002/2003	8	872	914.00	-1.01	
9	2003/2004	9	980	980.29	.00	
10	2004/2005	10	1047	1046.57	.00	
11	2005/2006	11	1113	1112.86	.00	
12						
13						
14						

Also this is the graph to best fit errors



4. Software Tools

4.1 SPSS

SPSS for Windows provides a powerful statistical analysis and data management system in a graphical environment, using descriptive menus and simple dialog boxes to do most of the work for you. Most tasks can be accomplished simply by pointing and clicking the mouse.

In addition to the simple point-and-click interface for statistical analysis, SPSS for Windows provides:

Data Editor: A versatile spreadsheet-like system for defining, entering, editing, and displaying data.

Viewer: The Viewer makes it easy to browse your results, selectively show and hide output, change the display order results, and move presentation-quality tables and charts between SPSS and other applications.

Multidimensional pivot tables: Your results come alive with multidimensional pivot tables. Explore your tables by rearranging rows, columns, and layers. Uncover important findings that can get lost in standard reports. Compare groups easily by splitting your table so that only one group is displayed at a time.

High-resolution graphics: High-resolution, full-color pie charts, bar charts, histograms, scatterplots, 3-D graphics, and more are included as standard features in SPSS.

Database access: Retrieve information from databases by using the Database Wizard instead of complicated SQL queries.

Data transformations: Transformation features help get your data ready for analysis. You can easily subset data, combine categories, add, aggregate, merge, split, and transpose files, and more.

Electronic distribution: Send e-mail reports to others with the click of a button, or export tables and charts in HTML format for Internet and intranet distribution.

Online Help: Detailed tutorials provide a comprehensive overview; context-sensitive Help topics in dialog boxes guide you through specific tasks; pop-up definitions in pivot table results explain statistical terms; the Statistics Coach helps you find the procedures that you need; and Case Studies provide hands-on examples of how to use statistical procedures and interpret the results.

Command language: Although most tasks can be accomplished with simple point-and-click gestures, SPSS also provides a powerful command language that allows you to save and automate many common tasks. The command language also provides some functionality not found in the menus and dialog boxes.

4.2 Recoding Variables

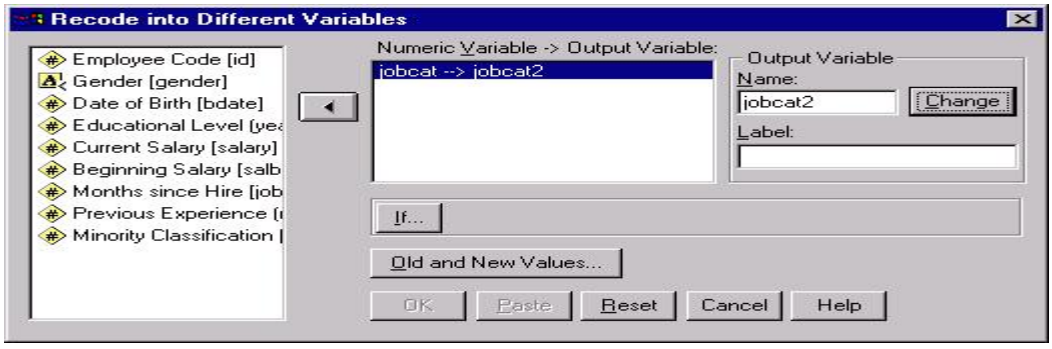
You can also modify the values of existing variables in your dataset. For example, if a dataset contains a variable that classifies an employee's status in three categories, but for a particular analysis you want to combine two of these classifications into a single category, then two of the values would need to be recoded into a single value so that there are two total groups.

The Recode option (or Alt+T+R) is available from the menu in the Data Editor:

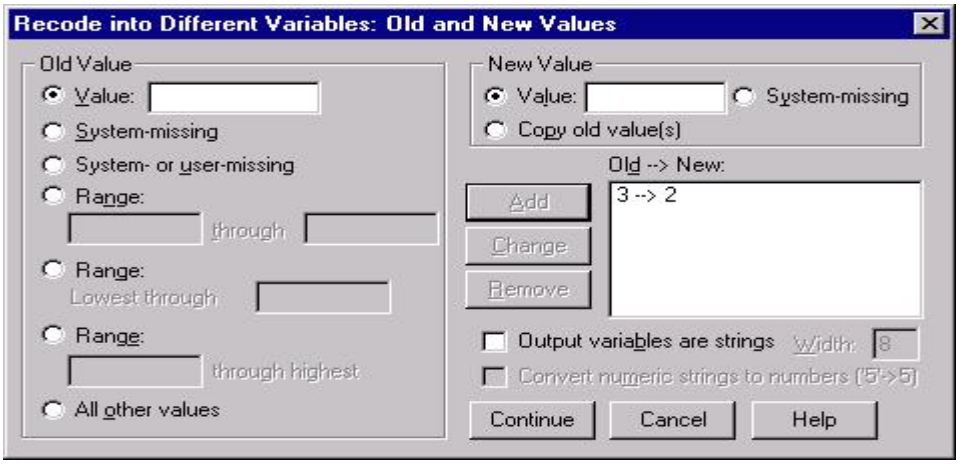
Transform
Recode

Additionally, there are two options for recoding variables in the Recode submenu. The Into Same Variables (Alt+T+R+S) option changes the values of the existing variables, whereas the Into Different Variables (Alt+T+R+D) option is used to create a new variable with the recoded values. Both options are essentially the same, except that recoding into a different variable requires you to supply a new variable name. You should use the Into Different Variables option, because you may change your mind about your recoding scheme at a later date. Thus, if your do change you mind, you still have the original values.

The following example illustrates the use of the Recode option to recode values into a new variable. When that option is selected from the menu, the following dialog box will appear:



First, a variable from the existing dataset should be selected by clicking on that variable, then clicking the arrow button in the middle of the dialog box. This will result in the selected variable being displayed in the box labeled, Numeric Variable -> Output Variable. Next, you must supply the name of the new variable, and optionally you can supply a label for the new variable. After a new variable name has been supplied, click on the button labeled Old and new Values. This will result in the following dialog box:



The above dialog box is the same regardless of whether you are recoding values into the same variable or creating a new variable. The original value of the variable being recoded is entered in the box labeled Old Value, and the new value is entered in the box labeled New Value. After values are entered in these boxes, click on the button labeled Add to complete the recode process.

Continuing with the above example, a variable with three values, such as jobcat, could be recoded into a variable with two values by recoding one of the values. In the example dataset, jobcat has three values: 1, 2, and 3. If the goal were to combine cases with the values 2 and 3, this could be accomplished by recoding cases with the value 3 into 2's. For example, by entering 3 in the box labeled Old Value and entering 2 in the box labeled New Value then clicking Add, all of the cases labeled 3 would take on the value 2. This can be repeated for as many of the values as necessary.

Values can also be recoded conditionally. The process for recoding values on the basis of a condition is essentially identical to the process for conditionally computing new variables discussed in the previous section: when you click on the If button in the main Recode dialog box, the same dialog box that was obtained from clicking If in the the Compute dialog box will appear with the same options.

4.3 Sorting Cases

Sorting cases allows you to organize rows of data in ascending or descending order on the basis of one or more variable. For example, the data could be sorted by job

category so that all of the cases coded as job category 1 appear first in the dataset, followed by all of the cases that are labeled 2 and 3 respectively. The data could also be sorted by more than one variable. For example, within job category, cases could be listed in order of their salary. The Sort Cases (or Alt+ D+O) option is available under the Data menu item in the Data Editor:

Data → Sort Cases...

The dialog box that results from selecting Sort Cases presents only a few options:

To choose whether the data are sorted in ascending or descending order, select the appropriate button. You must also specify on which variables the data are to be sorted. The hierarchy of such a sorting is determined by the order in which variables are entered in the Sort by box. Variables are sorted by the first variable entered, then the next variable is sorted within that first variable. For example, if jobcat was the first variable entered, followed by salary, the data would first be sorted by jobcat, then, within each of the job categories, data would be sorted by salary.

5. Conclusion

Regression models are powerful tools for predicting a score based on some other score. They involve a linear transformation of the predictor variable into the predicted variable. The parameters of the linear transformation are selected such that the least squares criterion is met, resulting in an "optimal" model. The model can then be used in the future to predict either exact scores, called point estimates, or intervals of scores, called interval estimates.

In this case study we find the optimal model in this formula:

$Y^{\wedge} = 383.714 + 66.286 * (X)$

Where Y^ is the predicted number of students

And X is the year we want to predict.

This model has only on dependent variable which is number of students in Salfeet Educational Region. Also we can develop this model to take multiple dependent variables like number of student in tawjihi in Salfeet Region and the population increasing percentage, but this required additional historical data to be built.

6. References

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[2] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

[3] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[4] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

A Stream-Based Selectivity Estimation Technique for Forward Xpath

muath alrammal and gaétan Hains
Université Paris Est, France
muath.alrammal@u-pec.fr , gaetan.hains@u-pec.fr

Abstract

Extensible Markup Language (XML) rapidly establishes itself as the de facto standard for presenting, storing, and exchanging data on the Internet. However, querying large volume of XML data represents a bottleneck for several computationally intensive applications. A fast and accurate selectivity estimation mechanism is of practical importance because selectivity estimation plays a fundamental role in XML query performance. Recently proposed techniques are all based on some forms of structure synopses that could be time-consuming to build and not effective for summarizing complex structure relationships. To overcome this limitation, we propose an innovative selectivity estimation algorithm, which consists of (1) the path tree synopsis data structure, a succinct description of the original document with low computational overhead and high accuracy for processing tasks like selectivity estimation, (2) the streaming selectivity estimation algorithm which is efficient for path tree traversal. Extensive experiments on both real and synthetic datasets show that our technique achieves better accuracy and less construction time than existing approaches.

Keywors: XML data, XPath queries, query optimization, stream processing, selectivity estimation.

[1] Introductin

XML [6] is currently being heavily pushed by the industry and community as the lingua franca for data representation and exchange on the Internet. The popularity of XML has created several important applications like information dissemination, processing of the scientific data, and real time news.
Query languages like XPath [4] and XQuery [5] have been proposed for accessing XML data. They provide a syntax for specifying which elements and attributes are sought to retrieve specific pieces of a document.
A stream of XML data is the depth-first, left-to-right traversal of an XML document [6]. Cost-based optimization of XML stream querying requires calculating the cost of XPath query operators. Usually the cost of an operator for a given XPath query depends heavily on the number of the final results returned by the query in question, and the number of temporary (intermediate) results that are buffered for its sub-queries [23]. Therefore, accurate selectivity estimation is necessary for cost-based optimization, but insufficient as we explain below.

Selectivity is a count of the number of matches for a query *Q* evaluated on an XML document *D*. This selectivity does not measure neither the size of these matches, nor the total amount of memory allocated for the temporary results. In addition, there are many parameters that influence streaming computational costs: the lazy vs eager strategy of the stack-automaton, the size and quantity of query results which depend on the query operator, the size and structure of the document , etc. The author of an XPath query may have no immediate idea of what to expect in

memory consumption and delay before collecting all the resulting sub-documents .

As a result, the current selectivity estimation techniques appear necessary but incomplete for managing queries on large documents accessed as streams. We therefore propose a new stream-based selectivity estimation technique. We compute the path tree, a synopsis data structure from the input XML document *D*. The purpose is to obtain a small but full structure synopsis that is traversed by an efficient streaming algorithm to reduce the computational overhead of complex XPath queries on *D*.

The remainder of the paper is structured as follows: the next section is a short survey of existing work on synopses data structures and twigs selectivity estimation. In the third section, we present our motivations and contributions. The fourth section presents our stream-based selectivity estimation technique. In the fifth section we compare our technique with the existing ones, and the paper then concludes with an outline of future work.

2. Related work

Various research works in estimating the selectivity of XPath queries have been published. The majority [1] [14] [13] [22] [11] have focused on linear XPath queries (e.g. //A//B/C). It is not clear how these approaches can be extended to XPath twig queries (quereis with predicates e.g. //A[.//B]/C) so as to cover a larger fragment of XPath.

Several structure synopses, such as Correlated Suffix Trees [7], Twig-Xsketch [17], TreeSketch [16], and XSeed [24], have been proposed for twig query selectivity estimation. They generally store some form of compressed tree structures and simple statistics such as node counts, child node counts, etc. Due to the loss of information, selectivity estimation heavily relies on the statistical assumptions of independence and uniformity. Consequently, they can suffer from poor accuracy when these assumptions are not valid. The above proposed structures synopses can not be evaluated by ordinary query evaluation algorithms, they require specialized estimation algorithms.

The authors of [7] proposed a correlated sub-path tree (CST), which is a pruned suffix tree (PST) with set hashing signatures that helps determine the correlation between branching paths when estimating the selectivity of twig queries. The CST method is off-line, handles twig queries, and supports substring queries on the leaf values. The CST is usually large in size and has been outperformed by [1] for simple path expressions.

Described in [17] the Twig-Xsketch is a complex synopsis data structure based on XSketch synopsis [13] augmented with edge distribution information. It was shown in [17] that Twig-Xsketch yields estimates with significantly smaller errors than correlated sub-path tree (CST). For the dataset XMark [18] the ratio of er- ror for CST is 26% vs. 3% for Twig-Xsketch.

TreeSketch[16] is found on a partitioned representation of nodes of the input graph-structured XML database. It extends the capabilities of XSketch [13] and [17] Twig-Xsketch. It introduces a novel concept of count-stability (C-stability) which is a refinement of the previous F-stability of [13]. This refinement leads to a better performance in the compression of the input graph-structured XML database.

Paper [15] introduced XCLUSTER, which computes a synopsis for a given XML document by summarizing both the structure and the content of document. The XCLUSTER-based synopsis data structure is a node- and edge-labelled graph, where each node represents a sub-set of elements with the same tag, and an edge connects two nodes if an element of the source node is the parent of elements of the target node. Nodes and edges of this graph are then equipped with special aggregate statistical information.

Paper [24] proposed the XSeed synopsis to summarize the structural information of XML data. The information is stored in two structures, a kernel, which summarizes the uniform information, and an HET (Hyper-Edge Table), which records the irregular information. By treating the structural information in a multi-layer manner, the XSeed synopsis is simpler and more accurate than the TreeSketch synopsis. Moreover, XSeed supports recursion by recording "recursion levels" and "recursive path expression" in the synopses. However, although the construction of XSeed is generally faster than that of TreeSketch, it is still time-consuming for complex datasets.

Paper [12] proposed a sampling method named

subtree sampling to built a representative sample of XML which preserves the tree structure and relationships of nodes. The number of data nodes for each tag name starting from the root level is examined. If it is sufficiently large, a desired fraction of data nodes are randomly selected using simple random sampling without replacement and the entire subtrees rooted at these selected data nodes are included as sampling units in the sample. If a tag has few data nodes at the level under study, then all the data nodes for that tag at the level are kept and they move down to check the next level in the tree. The path from the root to the selected subtrees are also included in the sample to preserve the relationships among the sample subtrees. Though a subtree sampling synopsis can be applied to aggregations functions such as SUM, AVG, etc., it is shown in [12] that XSeed [24] outperforms subtree sampling for queries with Parent/Child on simple dataset e.g. XMark [18], while it is the inverse for complex datasets.

3. Motivations and contributions

Having explored the state of the art, we summarize our motivations as follows:

- A 2005 study [20] of Yahoo's query logs revealed that 33% of the queries from the same user were repeated and that 87% of the time the user would click on the same result as earlier: repeat queries are used to revisit information [20]. This motivates our intense use of preprocessing: its cost can most often be amortized. Moreover it is possible to update our synopsis data structure by streamed and incremental updates.

- The proposed structures synopsis above (in section 2) can not be evaluated by ordinary query estimation structure, they require specialized estimation algorithms or rules.

- Though the construction time for structures synopsis vary, for example: the construction of XSeed is generally faster than that of TreeSketch as it is shown in [12]. The techniques used for synopsis construction are still time-consuming for complex datasets e.g. TreeBank [19].

- Most selectivity estimation techniques do not process the complete fragment of Forward XPath (the grammar of this frag ment is introduced in section 4.

Our contributions can be summarized as follows:

1. We present a new stream-based selectivity estimation technique. Where, we present the path tree, a synopsis structure for XML documents that is used for accurate selectivity estimates. We formally define it and we introduce a streaming algorithm to construct it. Furthermore, we introduce an efficient selectivity estimation algorithm for traversing the synopsis structure to calculate the estimates. The algorithm is well suited to be embedded in a cost-based optimizer.

2. Extensive experiments were performed. We considered the accuracy of the estimations, the types of queries and datasets that this synopsis can cover, the cost of the synopsis to be created, and the estimated vs measured memory allocated during query processing. Experiments demonstrated that out technique is both accurate and efficient.

4. Stream-based selectivity estimation technique

The stream-based selectivity estimation technique consists of (1) the path tree structure synopsis: a concise, accurate, and convenient summary of the structure of the XML document, (2) the selectivity estimation algorithm: an efficient streaming algorithm used to traverse the path tree synopsis to provide the end user with different estimates which allow him to optimize his query if needed.

The current version of our selectivity technique processes queries which belong to the fragment of Forward XPath*: a sub fragment of XPath 1.0 consisting of queries that have: child, descendant axis. NodeTest which is either element, wildchard, 'text()'. Predicate with ('or','not', 'and') and arithmetic operations.*

For a precise understanding of Forward XPath, we illustrate its grammar in figure 1. A location path is a structural pattern com- posed of sub expressions called steps. Each step consists of an axis (defines the tree-relationship between the selected nodes and the current node), a node-test (identifies a node within an axis), and zero or more predicates (to further refine the selected node-set). An absolute location path starts with a '/' or '//' and a relative location path starts with a './' or './/'. Where */node()* is a direct child, *//node()* is a descendant, *[./node()*] is a child predicate node for refinement, and *@node()* is an attribute;

```
Path := GenericPath
GenericPath := GenericStep | GenericStep GenericPath | GenericStep1
GenericStep := Axis NodeTest | Axis NodeTest '[' Predicate ']'
AttributeStep := '@' NodeTest  | '@' NodeTest '[' Predicate ']'
GenericStep1 := Axis NodeTest1
Axis := '/' | '//'
NodeTest := name | '*'
NodeTest1 := 'text()'
Predicate := PredicatePath | PredicatePath CompOp constant
                           | Predicate 'and' Predicate
                           | Predicate 'or' Predicate
                           | 'not(' Predicate ')'
CompOp := '=' | '!=' | '>' | '>=' | '<' | '<='
PredicatePath := '.' GenericPath | AttributeStep
```

Figure 1: Grammar of Forward XPath

Figure 2 illustrates our stream-based selectivity estimation technique. As shown in the figure, the path tree is built for the target XML document by using our streaming algorithm (explained in section 4.1.2). After that, the moment the end user sends an Xpath function estimator provides the end user with query's estimation by using the path tree and the selectivity estimation (explained in section 4.2).

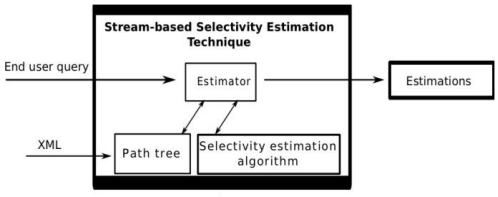


Figure 2: Stream-based selectivity estimation technique

Next, we will explain in details our technique.

4.1 Path tree

4.1.1 Path tree Definition

The path tree is a concise, accurate, and convenient summary of the structure of the XML dataset. It was invented by [1] but with a more restricted application than ours. To achieve conciseness, a path tree describes every distinct simple node-labelled path of a source XML exactly once with its frequency (the number of times it appears). To ensure accuracy, the path tree does not contain node-labelled paths that do not appear in the source XML dataset. The structure is convenient because it can be processed by ordinary query evaluation algorithms (stream-querying/stream-filtering algorithms) in place of the actual dataset.

Given an XML dataset D, the path tree is (a tree with node labels taken from D) defined as follows in figure 3. The details of path tree construction and updating are in [3]. However, we present below the pseudo code of the streaming algorithm for path tree construction with an example.

$\Sigma_{(D)}$ = is the finite set of nodes label of D.
$paths(D) = \{p = A_1, A_2, ...A_k \in \Sigma^*_{(D)} \mid p$ is a node-labelled path starting from the root of D *i.e.* A_1 is the root$\}$.
Remark: all node-label paths in $path(D)$ have A_1 as a prefix.

– **Definition:** we define $PathTree(D)$ as a graph whose nodes are $paths(D)$: $(path(D), \{(p_1, p_2) \mid \exists A \in \Sigma_{(D)}$ such that $p_2 = p_1 A$ *and* $p_i \in$ PathTree(D)$\}$).

– **Proposition:** $PathTree(D)$ is a tree rooted in root D.

– **Proof.** $T_{prefix} = (\Sigma^*_{(D)}, \{(p_1, p_2) \mid \exists A \in \Sigma_{(D)}$ such that $p_2 = p_1 A\})$ is the Hasse-diagram of the prefix relation on $\Sigma^*_{(D)}$ and has a tree structure. By construction $PathTree(D)$ is a subgraph of T_{prefix}. Therefore, $PathTree(D)$ is also a tree.

Figure 3: Path tree definition

4.1.2 Path tree Construction

To create a path tree from an XML dataset D, we consider that D is equivalent to a DFA and its path tree is equal to a minimized DFA. Minimization can be done by creating the DFA completely then applying the automata minimizing algorithm [10]. Another possibility which is more memory efficient is to generate the minimized DFA directly. In this paper, we propose a *streaming* algorithm which takes as input the SAX parser events of D and creates directly its minimized automaton. We explain our algorithm through the example below.

The minimized automata is illustrated in figure 4 (autoTable). We start by explaining the structure of this table. *nName*: is the label of the node, where $nName \in \Sigma_{(D)}$. *depth*: is the node's depthin *D*. *nDown* and *nUp*: are counters for naming the states in the automata (e.g. 1, 2, ...etc.). Their initialized values = 0. Note that $\delta(nDown, nName) = nUp$. *nFreq*: is the frequency of nName in *D* which have the same node-labelled path. *nSize*: is the size in byte of *nName* in *D* which have the same node-labelled path. A stack named pathStack is used to store the node-labelled path during the construction process of the path tree. At each SAX event *StartElement(nName)*, pathStack is pushed with *(nName, nDown)*, and at each *EndElement(nName)*, the top of pathStack is popped out.

When < A > the root of *D* is read, *depth* = 1 then, we add *A* with its information to accessAutoTable, autoTable and pathStack (algorithm 1 lines 2 − 7). Note that *nUp* of *A* =0. When < B > with *depth* = 2 is read, the function *checkSameNodePath* is called (algorithm 1 line 9). As long *B* is not yet a member of accessAutoTable (algorithm 2 line 1), then we add *B* with its information to accessAutoTable, autoTable and pathStack (algorithm 2 lines 21 − 27).

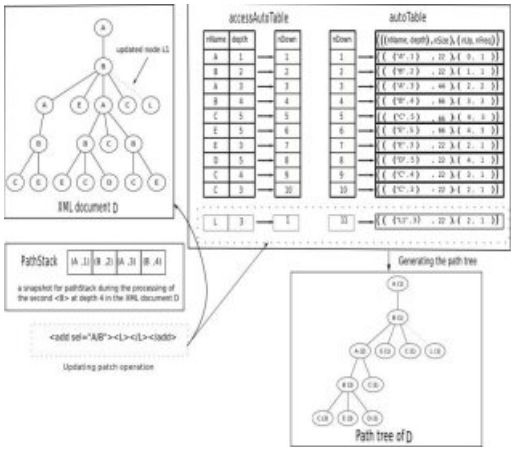


Figure 4: Path tree: Construction and Updating

The value of *nUp* for *B* with *depth* = 4 (which is already exist in autoTable) is 3 (see algorithm 2 line 5 and autoTable). Also, in pathStack the value *nDown* for the parent (*depth* −*1*) of the received *B* is 3 (see algorithm 2 line 6 and pathStack), both values are equals because the parents of both *nName B* have the same node-labelled path, which mean both *nName B* also have the same node-labelled path. Therefore, we increment the frequency and size of *B* (see algorithm 2 lines 8 − 11). If the node-labelled path of *B* was not exist in autoTable (see algorithm 2 line 12), then node *B* with its information is added (see algorithm 2 lines 13 − 19).

Algorithm 1: createAutoTable (*depth, nName, nSize*)

```
1 if (depth=1) then
2     nDown ← nDown + 1
3     nFreqStack= [1]          /*initialize the array with nFreq =1*/
4     nSizeStack= [nSize]      /*initialize the array with nSize (node
      size)*/
5     addNodeKey (depth, nName, nDown)          /*add a new node to
      accessAutoTable*/
6     addNode (nDown, nName, depth, nSizeStack, nUp, nFreqStack)  /*add
      a new node to autoTable.  Note that nUp = 0 */
7     pushPathStack (depth, nName, nDown)   /*update the pathStack*/
8 else
9     checkSameNodePath (depth, nName, nSize)
```

When the second < B > with *depth* = 4 is read, *B* is already a member of accessAutoTable (algorithm 2 line 1), therefore, we check whether the node-labelled path of the received *B* exists or not in autoTable (algorithm 2 lines 2 − 19).

The moment < /A > (EndElement of the root) is processed, the complete path tree can be generated and output in SAX events syntax.

The construction process is incremental, it allows constructing different incomplete path trees before the construction of the complete one. An *incomplete path tree* is the path tree for a part of an XML dataset.

Our streaming algorithm has time complexity $O(|depth(D)|.|D|)$ and space complexity $O(|depth(D)|.|pathTree(D)|)$.Where |D| is thr XML dataset size

Algorithm 2: checkSameNodePath (*depth, nName, nSize*)

```
1  if (isMemeber accessAutoTable (depth, nName)) then
2      l= get the list of all nDown in accessAutoTable which have the same key
       (depth, nName)
3      let nodePathExist = false
4      foreach nDown ∈ l do
5          nodenUp= get nUp of nDown from autoTable
6          nodenDownPathStack=get nDown of (depth − 1) from pathStack
7          if (nodenUp= nodenDownPathStack) then
8              nodePathExist= true
9              augmentFrequency (nFreqStack) /* augment the nFreq of
               nName by 1 */
10             augmentSize (nSizeStack, nSize) /* augment the value in
               nSizeStack by nSize */
11             pushPathStack(depth, nName, nDown)        /* update the
               pathStack */
12     if (isNodePathExist = false) then
13         nDown ← (nDown) + 1
14         nDownPathStack= get nDown of (depth − 1) from pathStack
15         nFreqStack= [1] /* initialize the array with nFreq=1  */
16         nSizeStack= [nSize]     /* initialize the array with nSize
           (node Size) */
17         addNodeKey (depth, nName, nDown)      /* add a new node to
           accessAutoTable */
18         addNode (nDown, nName, depth, nSizeStack, nDownPathStack,
           nFreqStack) /* add a new node to autoTable.  Note that
           nUp=nDownPathStack  */
19         pushPathStack (depth, nName, nDown)           /* update the
           pathStack */
20 else
21     nDown ← (nDown) + 1
22     nDownPathStack= get nDown of (depth − 1) from pathStack
23     nFreqStack= [1]     /* initialize the array with nFreq = 1  */
24     nSizeStack= [nSize]   /* initialize the array with nSize (node
       Size) */
25     addNodeKey (depth, nName, nDown)           /* add a new node to
       accessAutoTable */
26     addNode (nDown, nName, depth, nSizeStack, nDownPathStack,
       nFreqStack)      /* add a new node to autoTable.  Note that
       nUp = nDownPathStack  */
27     pushPathStack (depth, nName, nDown) /* update the pathStack */
```

Path tree updating: when the underlying XML dataset is updated, *i.e.* some elements are added or deleted, the path tree can be incrementally updated using XML patch operations [21]. Due to the space limitation, we explain this procedure by a short example. Figure 4 shown an example of a patch operation to update the XML dataset *D*. This operation adds an empty element *L* as a last child under *"A/B"* where element *A* is the root of *D*. The same patch will be sent to the path tree (accessAutoTable and autoTable) for updating. Thus, we check whether the node-labelled path of *L* that is *ABL* exists or not in autoTable. In this example, it is not, therefore we add the new node *L* with its information to accessAutoTable and autoTable (see figure 4). Otherwise (node-labeled path of *L* is exist), the frequency and the size of node *L* will be updated as we shown in algorithm 2 (lines 7-11).

4.2 Selectivity Estimation Algorithm

To enable the selectivity estimation process, we inspired our selectivity estimation algorithm from LQ (the extended lazy stream- querying algorithm of Gou and Chirkova work [9]). Therefore, the advantages of this algorithm are the same as for the lazy stream-querying algorithm. Detailed explanations about LQ and its advantages are in [2].

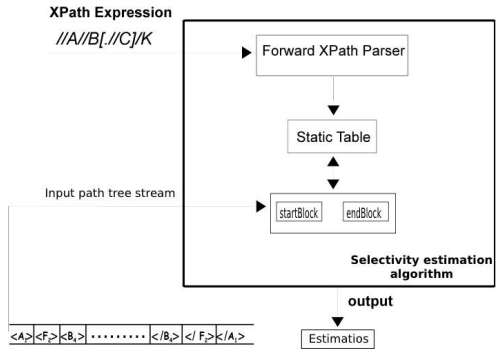


Figure 5: Selectivity estimation algorithm

Figure 5 illustrates our selectivity estimation algorithm. The current version of our estimation algorithm processes queries which belong to the fragment of Forward XPath. The estimation algorithm takes two input parameters. The first one is the XPath query that will be transformed to a query table statically using our Forward XPath Parser. After that, the main function is called. It reads the second parameter (the path tree) line by line repeatedly, each time generating a tag. Based on that tag a corresponding *startBlock* or *endBlock* function is called to process it. Finally, the main function generates as output the estimations needed for the given query.

Estimations are: *NumberOfMatches*: the number of answer ele- ments found during processing of the XPath query *Q* on the XML document *D*. *Cache*: the number of elements cached in the run-time stacks during processing of the XPath query Q on the XML document *D*. They correspond to the axis nodes of *Q*. *Buffer*: the number of potential answer elements buffered during processing of the XPath query *Q* on the XML document *D*. *OutputSize*: the total size in MiB of the number of answer elements found during processing of the XPath query *Q* on the XML document *D*. *WorkingSpace*: the total size in MiB for the number of elements cached in the run-time stacks and the number of potential answer elements buffered during processing of the XPath query *Q* on the XML document *D*.

NumberOfPredEvaluation: the number of times the query's predicates are evaluated (their values are changed or passed from an element to another).

The algorithms 3, 4, 5 and 6 are the pseudo code of the stack automaton (functions *startBlock* and *endBlock*) of our selectivity estimation algorithm. Detailed explanation of our algorithm and several examples on selectivity estimation process can be found in [2]. However, the pseudo code and the selectivity estimation process are explained through the example below.

4.2.1 Example on the Selectivity Estimation

Figure 6 illustrates different snapshots of the evaluation process of the path tree of *D* on the twig path //A[.//C]/B[.//D]//E which returns *E1*(3), *E2*(1) as result nodes. For each non-leaf node, the algorithm creates a stack. Therefore, in this example, a stack is created for the root node *A* and another one for the node *B*.

Algorithm 3: startBlock (*nName*, *nFreq*, *nSize*, *depth*)

```
1  if (parent stack of nNumber is not empty) then
2      if (node type ≠ Predicate) or (Predicate's value is still false) then
3          if (node axis =Descendant) or (node axis = Child) then
4              if (node = leaf) then
5                  if node type = Predicate) then
6                      evaluate the predicate node and increase the predicate
                       evaluation's counter (predCounter) by the value of
                       nFreq
7                  else
8                      if (node type =Result) then
9                          if (node is the query's root) then
10                             if (nName= text()) then
11                                 calculate: NumberOfMatches,
                                   OutputSize      /*Here we do not
                                   output the real value of the
                                   text node, in stead, we
                                   compute its real nSize and its
                                   nFreq */
12                                 output answers
13                             else
14                                 calculate: Buffer, WorkingSpace
15                                 buffer and append the node to the
                                   potential answers list of parent of the
                                   current node
16             else
17                 calculate: Cache, WorkingSpace
18                 push stack: nName, depth, list of the predicates, an empty
                   list for the potential answers, nFreq, nSize     /*the size
                   and the frequency of nName are pushed as
                   well.  */
```

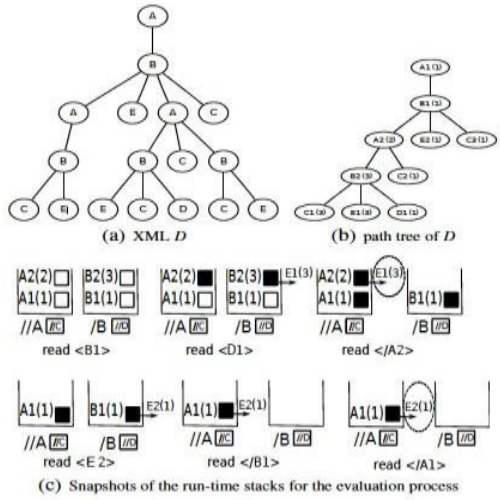


Figure 6: Snapshots of the run-time stacks for the evaluation of the path tree of D on Q (//A[.//C]/B[.//D]//E)

When < B2 > is read, the nodes *A1*(1), *B1*(1), and *A2*(1) were read and pushed (with their information) in their stacks. Concerning the node *B2*, it is also pushed (with its information) in its stack *B*. Note that for each pushed node, the values of *Cache* and *WorkingSpace* are updated. (algorithm 3 lines 16-17).

Algorithm 4: endBlock (*nName*, *nNumber*, *depth*)

```
1  if (node ≠ leaf) || (node's stack is empty) then
2      let s =get the top of the node's stack
3      if (node's depth = current depth) then
4          pop out the node
5          if (node's stack is not empty) then
6              check and update the predicates with descendant axis. If
               predicate node has a descendant axis, then increase predCounter
               by 1
7          let bool_Op_List= get the boolean operators associated with predicate
           children of the node
8          match (head bool_Op_List) with
9          | Not → if (the negation is true) then
10             processNodeType nNumber s              /*algorithm 5 */
11         else
12             appendOrDestroy nNumber s              /*algorithm 6*/
13         | And → if (all predicates are matched) then
14             processNodeType nNumber s     /*if the predicate does
               not contain a boolean operator, it will be
               processed as And.  */
15         else
16             appendOrDestroy nNumber s
17         | Or → if (one predicate is matched) then
18             processNodeType nNumber s
19         else
20             appendOrDestroy nNumber s
21         | Non → if (node has no predicate) then
22             processNodeType nNumber s
```

When < D1 > is read, the node *C1* was read, therefore, the value of the predicate C of *A2* was

changed to true, and the value of *NumberOf PredEvaluation* was updated. The node *E1* was read, therefore, *E1* was buffered (with its information) to the potential answers list of its parent node *B2*, and the values of *Buffer* and *WorkingSpace* were updated (algorithm 3 lines 13-14). Moreover, by reading *D1*, the value of the predicate *D* of *B2* was changed to true and the value of *NumberOfPredEvaluation* was updated.

Algorithm 5: processNodeType (*nNumber*, *s*)

```
1  if (node type = Axis) then
2      if (node is the query's root) then
3          let potential_answer_list = the list of the potential answers nodes of
           the current node
4          if (potential_answers_list of the current node is not empty) then
5              calculate: NumberOfMatches, OutpuSize
6              output the content of potential_answers_list: answers
7      else
8          if (potential_answers_list of the current node is not empty) then
9              append potential_answers_list to the same list of the parent of
               the current node
10 else
11     if (node type =Predicate) then
12         check and update the predicate and increase predCounter by 1
13         if (node axis = Descendant) then
14             clear the predicate's stack
15     else
16         if (node type =Result) then
17             if (node is the query's root) then
18                 calculate: NumberOfMatches, OutpuSize
19                 output answers
20             else
21                 append node to the potential answers list of the node's
                   parent
```

Algorithm 6: appendOrDestroy (*nNumber*, *s*)

```
1  if (node type = Axis) then
2      if the stack of the host node of the current node is empty then
3          destroy s
4      else
5          append the list of the potential answers of the current node to the same
           list of the top node of the host stack (the host stack of the current node)
```

When < /A2 > is read, the node *B2* was popped out from its stack, and the true value of its predicate *C* was passed to its ancestor *B1*, and the value of *NumberOfPredEvaluation* was updated (algorithm 4 line 6). Furthermore, the potential answers list of *B2* was appended to the same list of its parent node *A2* (algorithm 5 lines 8-9). Concerning *A2*, it is popped out of its stack, and as long as it is the root node, the content of its potential answers list is flushed as answers (algorithm 4 lines 13-14 then algorithm 5 lines 2-6).

When *E2* is read, it is buffered (with its information) to the potential answers list of its parent node *B2*, and the values of *Buffer* and *W orkingSpace* are updated (algorithm 3 lines 13-14). When < /B1 > is read, it is popped out from its stack and its potential answers list is appended to the same list of its parent node *A1*. Finally, when < /A1 > is read, it is popped out from its stack, *A1* is the root node, therefore, the content of its potential answers list is flushed as answers (algorithm 4 lines 13-14 then algorithm 5 lines 2-6).

The result of the XPath query estimation is as follows (estimated values): *NumberOfMatches*: the value is 4, they are: *E1*(3), *E2*(1) = 3 + 1 = 4. *Buffer*: in this example, the value of *Buffer* is the same as *NumberOfMatches*. *Cache*: the value is 7, we present them based on their stacks as follows: stack *A* contains *A1*(1), *A2*(2), while stack *B* contains *B1*(1), *B2*(3). The value then 1 + 2 + 1 + 3 = 7. *WoringSpace*: its size was estimated to 0.0002MiB. *OutputSize*: its size was estimated to 0.00008MiB.

The estimated values equal the real measured ones which shows the accuracy of our selectivity estimation technique.

5. Experimental Resutls

In this section, we demonstrate the accuracy of our technique by using variety of XML datasets and complex queries. Furthermore, we compare it with other approaches.

5.1 Experimental Setup

We performed experiments on a MacBook with the following technical specifications: Intel Core 2 Duo, 2.4 GHz, 4 GB RAM. The well known XML datasets XMark [18] and TreeBank [19] were selected for the experiments. XMark is a wide and shallow dataset, its size is 116MiB and its maximum depth is 12. TreeBank is a deep and recursive dataset, its size is 86MiB and its maximum depth is 36. The average relative error was used to measure the accuracy of our approach, it is defined as follows

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{M_i - P_i}{M_i}\right|$$

where M_i is the measured value of the $i-th$ query in the workload and P_i is its predicted one.

Extensive testing and complex Forward XPath queries were used in our experiments. For example, a complex XPath query applied to XMark

//item[.//payment or .//shipping]//mailbox//mail[./date] and to TreeBank //E MPTY [./S //N P[./∗] and .//V P]//∗/NNS.

5.2 Accuracy of selectivity estimation technique

Figure 7 illustrates the accuracy of our stream-based selectivity estimation technique

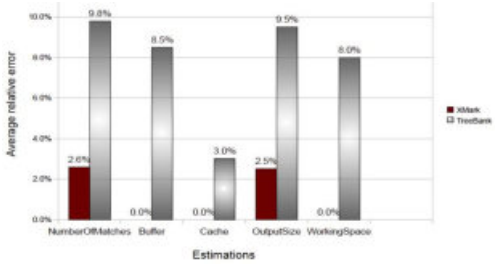


Figure 7: Accuracy of selectivity estimation technique

Our technique estimates the values of *NumberOfMatches*, *Cache*, *Buffer*, *OutputSize*, and *WorkingSpace*. While the different existing approaches estimate only the value of *NumberOfMatches*.

As shown in the figure, the accuracy of our technique on both datasets XMark and TreeBank is remarkable due to the complete structure information of the path tree which captures recursions in the dataset, and due to the efficiency of our selectivity estimation algorithm which supports the complete Forward XPath fragment. For example, in this figure, the max value of average relative error is for *NumberOfMatches* on TreeBank, it is less than 10%, so it is like informing the end user that the number of matches for an XPath query *Q* is 10 while in reality it is between 9 and 11.

5.3 Comparison with the existing techniques.

In this section, we compare our approach with other existing approaches.

5.3.1 Construction time for synopses

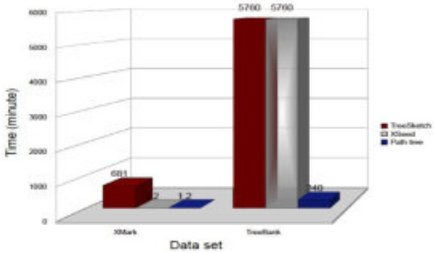


Figure 8: Construction time for synopses

Few approaches present the time needed to construct their synopses, like TreeSketch and XSeed. This is why in figure 8, we compare the construction time of our synopsis path tree with the same time needed for TreeSketch and Xseed.

Figure 8 shows the total construction time of TreeSktech, XSeed and path tree synopses. We do not show the construction time of the Subtree sampling synopsis because it is not a structural one (as we already explained in section 2), while for XCLUSTER and relational algebraic it is unknown.

The construction time of the structural synopses largely depends on the structure of the dataset. Our streaming algorithm for building path tree outperforms considerably the other approaches. The construction time for each of TreeSktech and XSeed for TreeBank 86MiB (depth 36) took more than 4 days (5760 minutes), this result was confirmed in [12]. While for path tree, the construction time for the same dataset took 244 minutes. Concerning XSeed, as the dataset become more complex, performance degrades dramatically [12] and construction time becomes significant. The construction time of path tree for TreeBank 86MiB (depth 36) is 24 times faster than XSeed (see figure 8).

5.3.2 Selectivity of structural queries: accuracy and synopsis size

TreeSketch and XSeed can estimate the accuracy for the number of matches (NumberOfMatches), while our approach estimates the accuracy for: *NumberOfMatches*, *Buffer*, *Cache*, and *OutputSize*, *WorkingSpace* . The accuracy of our approach outperforms the accuracy of TreeSketch and XSeed due to the complete structure of the path tree which captures the recursions in the dataset, and due to the efficiency our modified LQ algorithm which supports the complete Forward XPath fragment. The size of the path tree varies according to structure of the dataset. It is 10% of the size of TreeBank and 0.00006% of the size of XMark.

In all cases, an efficient streaming algorithm is used to traverse the path tree to avoid any computational overhead. Note that to control the space budget (synopsis size), it is possible to use a very partial, hence small path tree, to use no more space than competing approaches, but the accuracy of selectivity estimation will then be much lower.

The construction time for TreeSketch took more than 4 days. Actually we did stop the building process of its synopsis after 4 days. We faced the same situation for XSeed, but the difference between them that XSeed synopsis (as mentioned before) consists of two parts, an XSeed kernel and a hyper-edge table (HET). The kernel was built very fast, but the HET took more than 4 fays, this is why we did stop the construction process of HET. As long as, we could not build the synopsis of tree TreeSketch, and due to the fact that, the accuracy of XSeed outperforms the one of TreeSketch [24] [12], in figure 9 we compare our approach with the kernel of XSeed. We noticed that XSeed does not process queries with nested predicates and predicates with *'or', 'not', 'and'*, therefore, we refine and simplify the queries used in this experiment.

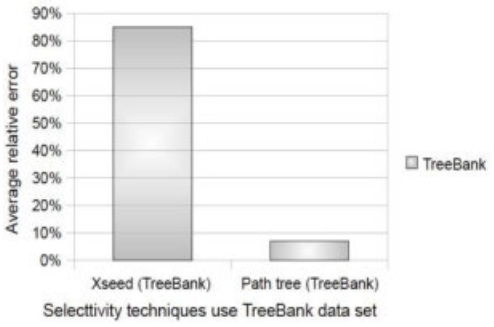


Figure 9: Accuracy of selectivity estimation techniques

As showed in figure 9, the average relative error of XSeed is almost 12 times higher than the same error for our approach.

5.3.3 Further Comparison

The fragment of XPath: the XPath fragment covered by our approach is more general than the one used by XSeed and TreeSketch. The TreeSketch query language does not support queries with *'text()'* [12] or with nested predicates. The XSeed query language does not support queries with *'text()'*, queries with nested predicates or queries with predicate which contain *'and', 'or', 'not'*.

Incremental update of synopsis: minimal synopsis size seems desirable but won't be the best because incremental maintenance would be difficult [8]. This is the case of TreeSketch and XSeed. While path tree preserves the same structure as the structure of its original XML dataset. So any language used to update the XML dataset can be used to update the path tree. Therefore, incremental update is possible, for example, by using the patch operations as we explained in section 4.1.2.

6. Conclusion and Perspectives

In this paper, we presented our stream-based selectivity estimation technique. It uses the path tree synopsis and an efficient selectivity estimation algorithm to provide the end user with different estimations which allow him to optimize his queries. Extensive experiments were performed to evaluate our technique. We considered the accuracy of estimations, the types of queries and datasets that the selectivity estimation technique can cover, and the cost of the synopsis to be created. Experiments demonstrated that our tech nique is accurate and outperforms the existing approaches.

As an undergoing research, we study how to compute a synopsis for a given XML dataset by summarizing both the structure and the content of the dataset

7. Acknowledgement

The authors thank M. Zergaoui president of Innovimax SARL for financial support in the form of a CIFRE scholarship for M. Alrammal, for suggesting the initial problem statement and participating in this work's supervision. Financial support from ANRT is also gratefully acknowledged

References

[1] A. Aboulnaga, A. R. Alameldeen, and J. F. Naughton. Estimating the Selectivity of XML Path Expressions for Internet Scale Applications. In Proc. Of the 27th (VLDB), pages 591 – 600, 2001.

[2] M. Alrammal. Algorithms for XML Stream Processing: Massive Data, External Memory and Scalable Performance. Thesis, Université Paris-Est, 2011. http://lacl.univ=paris12.fr/Rapports/TR/muth_thesis.pdf.

[3] M. Alrammal, G. Hains, and M. Zergaoui. Path tree: Document Synopsis for XPath Query Selectivity Estimation. IEEE, In Proc. of the 5th (CISIS), pages 321–328, 2011.

[4] A. Berglund, S. Boag, D. Chamberlin, M. F. Fernández, M. Kay, J. Robie, and J. Siméon. XML Path Language (XPath) 2.0. 14 December 2010. *http://www.w3.org/TR/2010/REC-xpath20-20101214/*.

[5] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, and J. Siméon. XQuery 1.0: An XML Query Language (Second Edition). 14 December 2010. *http://www.w3.org/TR/2010/REC-xquery-20101214/*.

[6] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and F. Yergeau. Extensible Markup Language (XML) 1.0 (fifth edition). 26 November 2008. *http://www.w3.org/TR/REC-xml/*.

[7] Z. Chen, H. V. Jagadish, F. Korn, N. Koudas, S. Muthukrishnan, R. T. Ng, and D. Srivastava. Counting

Twig Matches in a Tree. In Proc. of the 17th (ICDE), pages
595 – 604, 2001.

[8] R. Goldman and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In Proc. of the 23rd (VLDB), pages 436–445, August 1997.

[9] G. Gou and R. Chirkova. Efficient Algorithms for Evaluating XPath over Streams. In Proc. of the 2007 ACM SIGMOD, pages 269–280, 2007.

[10] J. Hopcroft and J. Ullman. Introduction to Automata Theory, Language, and Computation. 1979.

[11] C. Luo, Z. Jiang, W.-C. Hou, F. Yan, and C.-F. Wang. Estimating XML Structural Join Size Quickly andEconomically. In Proc. of the 22nd (ICDE), 2006.

[12] C. Luo, Z. Jiang, W.-C. Hou, F. Yu, and Q. Zhu. A Sampling Approach for XML Query Selectivity Estimation. In Proc. of the (EDBT), pages 335–344, 2009.

[13] N. Polyzotis and M. Garofalakis. Statistical Synopses for Graph-structured XML Databases. In Proc. of the 2002 ACM SIGMOD, pages 358–369, 2002.

[14] N. Polyzotis and M. Garofalakis. Structure and Value Synopses for XML Data Graphs. In Proc. of the 28th (VLDB), pages 466–477, 2002.

[15] N. Polyzotis and M. N. Garofalakis. XCluster Synopses for Structured XML Content. In Proc. of (ICDE), 2006.

[16] N. Polyzotis, M. N. Garofalakis, and Y. Ioannidis.Approximate XML Query Answers. In Proc. of the 2004 ACM SIGMOD, pages 263–274, 2004.

[17] N. Polyzotis, M. N. Garofalakis, and Y. Ioannidis. Selectivity Estimation for XML Twigs. In Proc. of the (ICDE), 2004.

[18] A. Schmidt, R. Busse, M. Carey, M. K. D. Florescu, I. Manolescu, and F. Waas. Xmark: An XML Benchmark Project. Technical report, 2001. *http://www.xml-benchmark.org/*.

[19] D. Suciu. Treebank: XML Data Repository. Technical report, University of Pennsylvania Treebank Project, Novmber 1992. *http://www.cs.washington.edu/research/xmldatasets*.

[20] J. Teevan, E. Adar, R. Jones, and M. Potts. History Repeats Itself: Repeat Queries in Yahoo's Query Logs. In Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 703–704, 2005.

[21] J. Urpalainen. XML Patch Operations Framework Utilizing XPath Selectors. Network Working Group, 2008. *http://datatracker.ietf.org/doc/rfc5261/*.

[22] W. Wang, H. Jiang, H. Lu, and J. X.Yu. Containment Join Size Estimation: Models and Methods. In Proc. of the 2002 ACM SIGMOD, pages 145–156, 2003.

[23] N. Zhang, P. Haas, V. Josifovski, G. Lohman, and C. Zhang. Statistical Learning Techniques for Costing XML Queries. In Proc. of the 31st (VLDB), pages 289–300, 2005.

[24] N. Zhang, M. T. Ozsu, A. Aboulnaga, and I. F. Ilyas. XSeed: Accurate and Fast Cardinality Estimation for XPath Queries. In Proc. of the 20th (ICDE), 2006

Comparison Study of Adhoc Networks Routing Protocols Using NS2

Ola Sbihat,
Arab Amirican University, Palestinian
osbihat@hotmail.com

Abstract:

Adhoc networks are the next step in the evolution of wireless architecture, delivering wireless services for a large variety of applications; they are very useful in emergency search-and-rescue operations, in the applications where the persons wish to quickly share information, and data acquisition operations in inhospitable terrain. Ad hoc wireless networks are increasingly gaining importance due to their advantages such as low cost and ease of deployment . In recent years, a variety of new routing protocols targeted specifically at this environment have been developed like Destination-Sequenced Distance-Vector (DSDV), Dynamic Source Routing (DSR), Ad Hoc On-Demand Distance Vector Routing (AODV). In this report, I will present a comparison study of the performance of these routing algorithms in ad hoc networks under the IEEE 802.11s specifications

1. Introduction

Ad hoc network consists of nodes which are mobile and can be connected in an arbitrary manner. All nodes of these networks behave as routers. The term routing refers to the process of selecting paths in a computer network along which to send data. This process can be defined as a routing protocol, used to exchange information about topology and link weights, computes paths between nodes and all of that can be done using a routing algorithm.

Mobile ad-hoc networks (MANETs) are self-organized networks. Communication in an ad-hoc network does not require existence of a central base station or a fixed network topology. Each node of an ad-hoc network can be both a host and a router. As well as destination of some information packets while at the same time it can act as relay station for other packets to get their final destination. This makes communication between nodes outside direct radio range of each other possible, is probably the most distinct difference between mobile ad-hoc networks and wireless LANs [1].

Traditional routing protocols cannot perform in such environment resulting in development such routing protocols for ad hoc networks, i.e. AODV, DSR and DSDV.

The rest of the report is organized as follows; in the following section, I will briefly review AODV, DSDV and DSR protocols. In next section; I

will present detailed observation on simulation environment. Finally, presents Simulation results, analysis followed by conclusions.

2. Description of the Ad-hoc Routing Protocols

2.1. Ad hoc On Demand Distance Vector (AODV)

AODV routing protocol designed for ad hoc mobile networks and it is suitable for both unicast and multicast routing [1]. The meaning of demand that it builds routes between nodes only as preferred by source nodes. It maintains these routes as long as they are needed by the sources. Moreover, AODV forms trees which connect multicast group members. The trees are composed of the group members and the nodes needed to connect the members. AODV uses sequence numbers to ensure the freshness of routes. It is loop-free, self-starting, and scales to large numbers of mobile nodes [2].

AODV builds routes using a route request / route reply query cycle. The source node broadcasts a route request (RREQ) packet across the network when its needs a route to a certain destination which it doesn't have rout information about it. The other nodes receiving this packet update their information for the source node and set up backwards pointers to the source node in the route tables and the source node's IP address, current sequence number, and broadcast ID, the RREQ also contains the most recent sequence number for the destination of which the source node is aware. A node receiving the RREQ may send a route reply (RREP) if it is either the destination or if it has a route to the destination with corresponding sequence number greater than or equal to that contained in the RREQ. If this is the case, it unicasts a RREP back to the source [3]. Otherwise, it rebroadcasts the RREQ. Nodes keep track of the RREQ's source IP address and broadcast ID. If they receive a RREQ which they have already processed, they discard the RREQ and do not forward it. As the RREP propagates back to the source, nodes set up forward pointers to the destination. Once the source node receives the RREP, it may begin to forward data packets to the destination. If the source later receives a RREP containing a greater sequence number or contains the same sequence number with a smaller hop count, it may update its routing information for that destination and begin using the better route.

Nodes monitor the link status of next hops in active routes, when a link break in an active route is detected; a RERR message is used to notify other nodes that the loss of that link has occurred. The RERR message

indicates which destinations are now unreachable due to the loss of the link.

2.2. Destination Sequenced Distance Vector (DSDV)

DSDV is a Table driven routing protocol and uses sequence numbers to mark each node to improve upon the loop problem. Routing information is distributed between nodes via sending "full dumps" and incremental updates. "Settling time" [4] metric is used to determine update interval. Each node maintains a routing table consisting of entries with each for a destination. Each entry contains a metric to that destination and the recently sequence number broadcast from that destination. Upon receiving an update from a neighbor, a node updates an entry in its own routing table if, for that entry, the update contains a higher sequence number or the update contains a same sequence number but a shorter metric than that has been seen before.

To update an entry, a node sets the metric in its table entry for that destination to one hop more than the metric in that neighbor's update. When a node sends an update message, it puts a sequence number in the entry for itself in that update and sets the metric value to zero; for each of other entries, it duplicates all the entries maintained in its own routing table.

Clearly, the sequence numbers and metric values containing in each update play a vital role in DSDV operation. A malicious node can easily disrupt the routing protocol by arbitrarily tempering the sequence numbers or the metrics [4] [5].

2.3. Dynamic Source Routing (DSR)

The key distinguishing feature of DSR is the use of source routing. That is, the sender knows the complete hop-by-hop route to the destination. These routes are stored in a route cache. The data packets carry the source route in the packet header. When a node in the ad hoc network attempts to send a data packet to a destination for which it does not already know the route, it uses a route discovery process to dynamically determine such a route. Route discovery works by flooding the network with route request (RREQ) packets. Each node receiving an RREQ rebroadcasts it, unless it is the destination or it has a route to the destination in its route cache. Such a node replies to the RREQ with a route reply (RREP) packet that is routed back to the original source. RREQ and RREP packets are also

source routed. The RREQ builds up the path traversed across the network. The RREP routes itself back to the source by traversing this path backward. The route carried back by the RREP packet is cached at the source for future use. If any link on a source route is broken, the source node is notified using a route error (RERR) packet. The source removes any route using this link from its cache. A new route discovery process must be initiated by the source if this route is still needed. DSR makes very aggressive use of source routing and route caching. No special mechanism to detect routing loops is needed. Also, any forwarding node caches the source route in a packet it forwards for possible future use.

3. simulation parameters

3.1. Performance Metrics

The comparison between the three routing protocols in this projects concentrate on 3 performance metrics which are:

Throughput: is the ratio of total amounts of data that reaches the receiver from the source to the time taken by the receiver to receive the last packet. It is represented in packets per second or bits per second. In the Ad Hoc Networks unreliable communication, limited energy, limited bandwidth and frequent topology change affect throughput [4][6].

Average end-to-end delay of data packets: the average time that a packet takes to traverse the network. This is the time from the generation of the packet in the sender up to its reception at the destination's application layer and it is measured in seconds. It therefore includes all the delays in the network such as buffer queues, transmission time and delays induced by routing activities and MAC control exchanges. Various applications require different levels of packet delay. Delay sensitive applications such as voice require a low average delay in the network whereas other applications such as FTP may be tolerant to delays up to a certain level. Ad hoc Networks are characterized by node mobility, packet retransmissions due to weak signal strengths between nodes cause the delay in the network to increase. The End-to-End delay is therefore a measure of how well a routing protocol adapts to the various constraints in the network and represents the reliability of the routing protocol [7] [8].

Average Routing Overhead: is the total number of routing packets divided by total number of delivered data packets. This metric provides an

indication of the extra bandwidth consumed by overhead to deliver data traffic. It is crucial as the size of routing packets may vary. The routing overhead describes how many routing packets for route discovery and route maintenance need to be sent in order to propagate the CBR packets. It is an important measure for the scalability of a protocol. It for instance determines, if a protocol will function in congested or low-bandwidth situations or how much node battery power it consumes [6].

3.2. The Simulation Assumptions

The simulations in this report made using the Network Simulator (NS-2) which provides the implementation of the DSR, AODV, and DSDV. I run the simulation in accept as in the later on described scenario. The detailed trace file created by each run is stored to disk, and analyzed using a variety of scripts "*.tr"* files for example the file that counts the number of packets successfully delivered and the length of the paths taken by the packets, as well as additional information about the internal functioning of each scripts executed. In each time I collected these data from the trace file and made the simulations many times (five times exactly). I stored the results in a MATLAB file (a matrix) and I took the average of the results and plot these results using MATLAB.

The following assumptions are considered when building the Tcl script:

- *Traffic models*

Random traffic connections of TCP and CBR can be setup between mobile nodes using a traffic-scenario generator script. This traffic generator script is available under ~ns/indep-utils/cmu-scen-gen and is called cbrgen.tcl. It can be used to create CBR and TCP traffics connections between wireless mobile nodes. For the simulations carried out, traffic models were generated for 40 nodes with cbr traffic sources.

- *Mobility models*

The node-movement generator is available under ~ns/indep-utils/cmu-scen-gen/setdest directory. Mobility models were created for the simulations, with simulation times of 0,50,100,150,200 seconds, maximum speed of 20m/s, topology boundary of 1500x300 and simulation pause time of 60secs.

All flows in the system are assumed to have the same type of traffic source. All the senders have traffic with the rate of data rate/number of stations packet per second; in summary; Table 1 shows the simulation parameters used in the simulations.

Table 9 : Simulation Parameters

Parameter	*Value*
Simulation time	0,50,100,150, 200 s
Routing Protocol	DSR, AODV & DSDV
Pause Time	60 s
Mac Type	802.11
Number of Nodes	10,25,40,75, 100
Environment Size	1500 X 300
Speed	20 m/s
Traffic Type	CBR,TCP
Packet Size	512 Bytes

4. RESULTS AND PERFORMANCE

Performance of AODV, DSR and DSDV protocols is evaluated under both CBR and TCP traffic pattern. Simulation is done by using NS-2. Varying the simulation time and Number of Sources to see the performance difference between these protocols for each performance metric parameter.

4.1. Throughput:

The unit of throughput is Mbps, however we have taken Kilo bits per second (Kbps). Varying the number of nodes and obtaining the throughput values for each number of nodes in the CBR traffic shown in Figure 1.

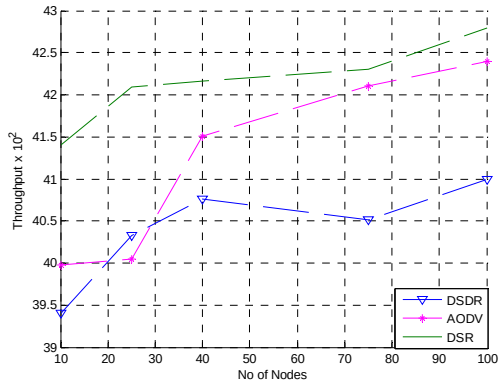


Figure 1: Number of Nodes vs. Throughput for CBR traffic

Now, the simulations repeated for 40 nodes and with various simulations time shown in Table 1. The packets were sent at a rate of 10 packets/sec. Figure 2 describes the results.

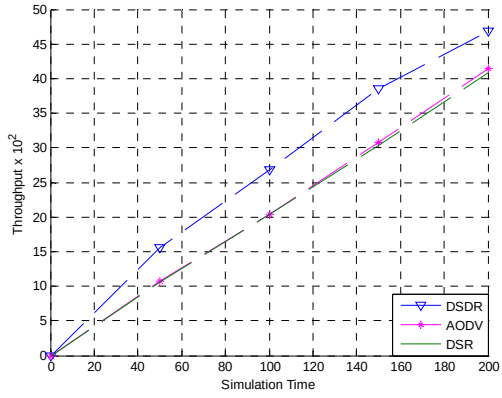


Figure 2: Simulation Time vs. Throughput for CBR traffic

In case of TCP traffic, throughput increases in slow amount for all three protocols independent of number of nodes as shown in Figure 3. While

throughput changes rapidly when varying the simulation time as cleared in Figure 4.

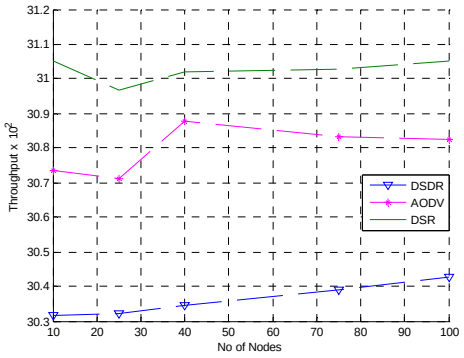


Figure 3: Number of Nodes vs. Throughput for TCP traffic

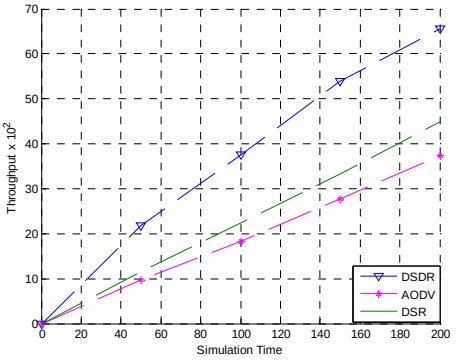


Figure 4: Simulation Time vs. Throughput for TCP traffic

4.2. Average End to End Delay Result

When the buffers become full quickly in CBR traffic, so the packets have to stay in the buffers longer period of time before they are sent. For average end-to-end delay, the performance of DSR decreases and varies with the number of nodes. However, the performance of DSDV is degrading due to increase in the number of nodes the load of exchange of routing tables becomes high and the frequency of exchange also increases due to the mobility of nodes. The performance of AODV increases and remains constant as the number of nodes increases as in Figure 5.

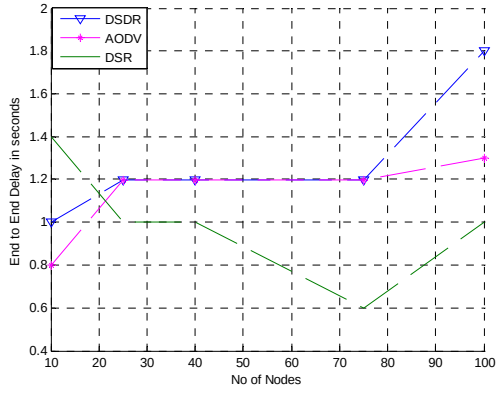


Figure 5: Average End-to-End Delay for Vs No. of Nodes in CBR Traffic

In TCP as in Figure 6. Average end-to-end Delay is also remains almost constant in DSDV while it varies in the case of AODV and DSR protocols with respect to change in simulation time.

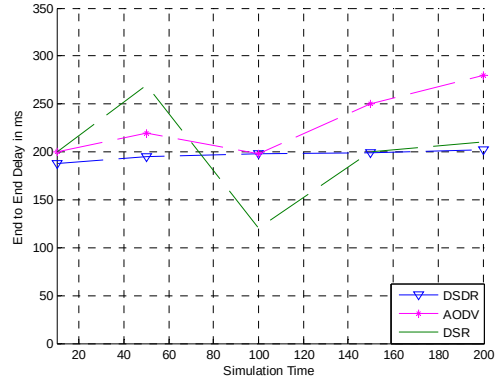


Figure 6: Average End-to-End Delay for Vs Simulation Time in CBR Traffic

In case of TCP Figures7show the average End-to-End delay for the DSDV, AODV and DSR protocols for various numbers of nodes.

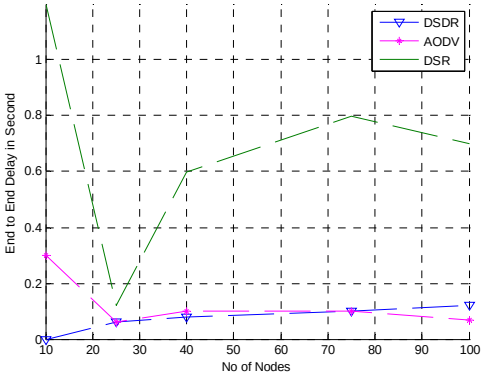


Figure 7: Average End-to-End Delay for Vs. No. of Nodes in TCP Traffic

It is clear that in Figure 8 that DSDV has the shortest End-to-End delay than AODV and DSR since DSDV is a proactive protocol and all routing information are already stored in table. Hence, it consumes lesser time.

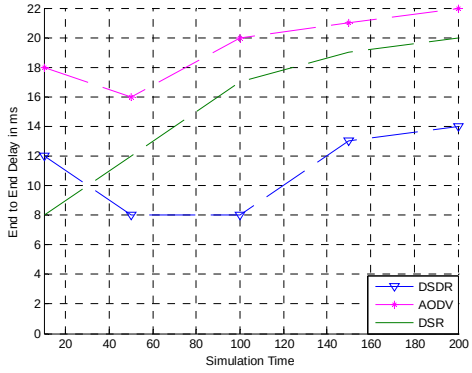


Figure 8: Average End-to-End Delay for Vs Simulation Time in TCP Traffic

4.3. Average Routing Overhead Results

Figure 9 illustrate the performance for average routing overhead required by all three protocols when subjected to various numbers of nodes. This metric gives an idea of the extra bandwidth that is required to deliver the data packets.

It can be seen that DSR exhibits the highest average routing overhead because of its route cache property. It generates the highest no. of routing packets but its loss of packets is also more. Moreover, AODV routing overhead gradually increases in case change in no. of nodes.

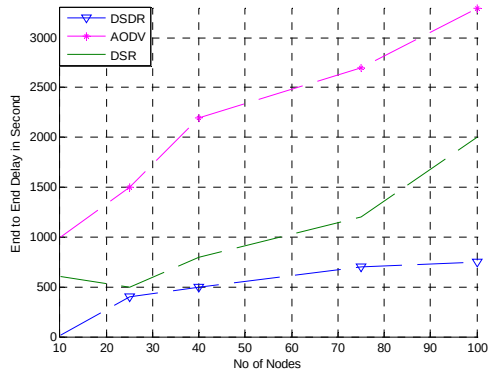


Figure 9: Average Routing Overhead Vs No of Nodes

However, AODV starts decreasing as simulation time is increased as in Figure 10. DSDV is independent of change in simulation time and no. of nodes. Routing overhead is lowest and constant in both test cases

because of its table-driven nature. However, it gradually increases a bit for change in no. of nodes.

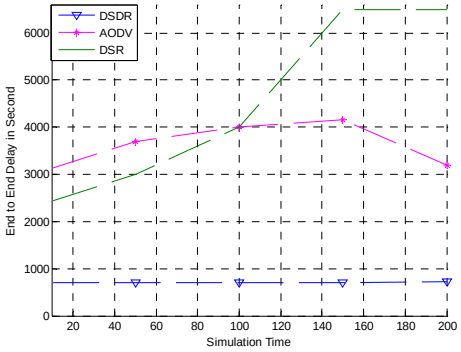


Figure 10: Average Routing Overhead Vs Simulation Time

5. Conclusions

This project compared the performance of DSDV, AODV and DSR routing protocols for ad hoc networks using ns-2 simulations based on both CBR and TCP traffic. These routing protocols were compared in terms of Throughput, Average end-to end delay and Average Routing Overhead when subjected to change in no. of nodes and the simulation time. Simulation results show:

- DSR shows higher throughput than the DSDV and AODV since its routing overhead is less than others. The rate of packet received for AODV is better than the DSDV.
- End-to-end delay in AODV is not affected by change in simulation time. It is affected when no. of nodes is changed; however, it gets stable as the no. of nodes is increased. Its performance is similarity to DSDV.
- DSDV performs better than DSR and AODV as far as average end-to-end delay is concerned. End-to-end delay in DSDV is independent of any change in simulation time or no. of nodes. It is lowest and most stable in both test cases.
- In terms of average routing overhead DSDV performs better than AODV and DSR. AODV follows DSDV closely for average routing overhead.
- AODV performance is the best considering its ability to maintain connection by periodic exchange of information, which is required for TCP, based traffic.
- It is also true that any of the three protocols is the best. Their performance depends upon the different scenarios.

6. References

[1] J. chen, y. z. Lee & M. perla "Performance comparison of AODV and OFLSR in wireless networks"

[2] http://moment.cs.ucsb.edu/AODV/aodv.html

[3] http://www.crhc.illinois.edu/wireless/assignments/simulations/lab109.html

[4] T. Liu & K. Liu, "Improvement on DSDV in Mobile Ad Hoc Networks", IEEE, China, 2007

[5] D. B. Johnson, D. A. Maltz & J. Broch. "DSR: The Dynamic Source Routing protocol for Multi-Hop Wireless Ad Hoc networks".

[6] A. Abdullah, N. Ramly, A. Muhammed & M. N. Derahman."Performance Comparison Study of Routing Protocols for Mobile Grid Environment", IJCSNS

International Journal of Computer Science and Network Security, Vol.8, 2008

[7] U. R. Khan, K. Reddy, A. V. Zaman, R. U. Reddy & K.A. Harsha, "An efficient DSDV routing protocol for MANET and its usefulness for providing Internet access to Ad Hoc Hosts, IEEE, Nov. 2008.

[8] S. R. Das, R. Castaneda & J. Yan, "Simulation based performance evaluation of routing protocols for mobile ad hoc networks", http://www.cib.espol.edu.ec/digipath/d_papers/ 40518.pdf

[9] Ns2 tutorial- http://www.isi.edu/nsnam/ns/

[10] http://www.isi.edu/nsnam/ns/ns-tutorial/index.html

[11] S. Shah, A. Khandre, M. Shirole & G. Bhole, "Performance Evaluation of Ad Hoc Routing Protocols Using NS2 Simulation". Int. J. Advanced Networking and Applications, Vol. 2, 2011.

Possibility of Applying Green Communications in Palestinian Cellular Networks

Murad Abusubaih, Yaqoub Sharabati, Omar Maraqa, Sayf Najm Eddin
Department of Electrical and Computer Engineering
Palestine Polytechnic University
Hebron, Palestine
murads@ieee.org

Abstract:

Due to the large deployment of wireless networks, energy efficient networks have attracted the attention of researches in recent years. The main challenging aspect is the development of policies and protocols for enabling energy saving. Green Communication is a key idea in this direction. It mainly focuses on network design that enables activation of resources on demand. This paper studies the possibility of applying the green communication idea in Palestinian cellular networks. Detailed simulation experiments have been conducted using real traces from Palestinian cellular operators. We present simulation results that demonstrate the amount of energy saving when green communication is applied. Further, we demonstrate the effect of green communication on Quality of Service experienced by cellular users.

Keywords: *Green Communication, Cellular, Networks, energy consumption, Optimization.*

1. Introduction

Recently, the number of cellular subscribers has reached 4 billion [1]. 120,000 new base stations (BS) are being deployed every year to serve 400 million new mobile subscribers around the world [2].

The huge number of BSs, coupled with the large number of cellular mobiles consume tremendous amount of energy. This situation has triggered the attention of network designers and researchers to seek for energy efficient networks to reduce energy costs. Towards this goal, a new innovative research field has been created under the name of ***"Green Communications"***.

Green Communication includes policies and protocols that enable a reduction of the energy consumed by the network entities. Most efforts in

this direction are devoted to design of optimal energy saving approaches at BSs.

It is known that cellular operators are primarily focusing on technological developments that meet consumers' capacity and Quality of Service (QoS) requirements as well as increased broadband data rates in order to support real time applications. However, the increasing awareness of environment and economic together with the high cost of energy have posed challenges on improving power efficiency in communication systems. In [3], it has been pointed out that information and communication technologies consume about 2% to 10% of the world power consumption, where cellular networks consume the most. BSs together with their supporting equipments consume 80% of energy consumed by a cellular network [4]. Each BS consumes 35MWh per year [2]. Therefore, a special focus is being given to reduction of energy consumed by these units. This can be achieved by switching off some BSs or some cells (sectors) within a BS during low traffic periods.

In this work, we investigate the possibility of applying the green communication concept within Palestinian cellular networks. Our study is based on real traces regarding traffic and network topologies obtained from operators in the country. The rest of this paper is organized as follows: Section 2 discusses the related work. In section 3, details of green communication and insights on the paper work are provided. Experimental setup is discussed in section 4. Simulation results are provided in section 5, before we conclude the paper in section 6.

2. Related Work

Reduction of energy consumption in wireless networks has significantly attracted the attention of several research groups in recent years. A focus is given to energy efficient strategies for wireless sensor network (WSN), Wireless Local Area Networks (WLAN) and cellular networks, including 2G, 3G and 4G.

In [1], Marsan et. al evaluate the amount of energy that can be saved by using two networks under high traffic conditions. They switched off one network during low traffic periods, assuring at the same time QoS obtained when both networks are active. The authors demonstrate a 25%-35% energy saving. This efficiency is attributed to the fact that

network operators plan their networks based on peak times.

The authors of [3] focus on the dynamic planning of UMTS networks. They consider switching off some UMTS cells and Node B's in urban areas during low traffic periods, while still guaranteeing QoS constraints in terms of blocking probability. The authors conclude that in some scenarios, it is possible to reduce power consumption of the network by 50%.

The authors of [5] provide simple analytical models that study the energy-aware management of cellular access networks. They try to characterize the amount of energy that can be saved by reducing the number of active cells during low traffic periods. The paper assumes that traffic is uniform across cells and neighbouring cells fill the coverage of switched off cells. The paper encourages the cellular network operators to consider devised approaches for dynamic management of network resources, so as to obtain very large energy savings.

Several extensions are also suggested including the consideration of cell breathing, antenna tilting and power control.

The papers of [6] and [7] address power saving in WLANs. They provide strategies for reducing power consumption within WLANs without affecting clients' performance.

In [6], the authors are concerned about high density WLANs, wherein the possibility of applying green communication may be affordable, by switching off some Access Points (APs) during low traffic periods.

Reference [7] develops improved analytical model used for simple on-demand polices in dense WLANs. The authors concentrate on saving the energy of underutilized APs. Such APs are candidates to be switched off. Their users are handed off to other ones.

These encouraging results have triggered us to study the potential benefits of applying the green communication concept within Palestinian cellular networks, where two operators are currently active and a licence for a third one is under consideration.

3. Applying Green Communication in Palestinian Networks

In this work, we study the possibility of applying the green communication in Palestinian cellular networks. We

consider two cellular network operators, which we refer hereafter as Operator 1 and Operator 2. Each BS operates 3 cells (sectors). We concentrate on three cases. In the first, we try to switch off some cells within one operator network, while in the second we switch off all cells within a network owned by one operator and roam potential active users to the second network. In the third case, collaboration among BSs of both operators is assumed and we switch off some cells in both networks and roam potential active users to the best cells regardless of their operator. The selection strategy of cells to be switched off is based on traffic conditions and possibility of roaming users to other cells assuring acceptable level of Signal to Noise Ratio (SNR). In each case, we compute the percentage of energy saving as well as the number of users that experience an SNR below a target threshold due to roaming.

The power saved in one day is calculated as follows:

$P_{res} = (P_{con} * T * C_{off}) \rightarrow (1)$

Where:

P_{res}: is the amount of power saved in KW.

P_{con}: Power consumed from a cell per hour in KW.

T: Time of cells turned off in hours.

C_{off}: # of cells turned off.

The motivation of the work stems from our observation of the traffic profiles obtained from both operators. They are shown in figures 1 and 2.

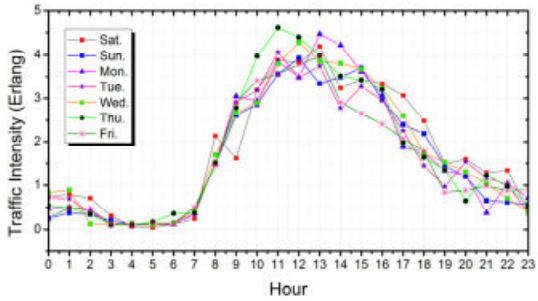


Fig. 1 Traffic Profile of Operator 1

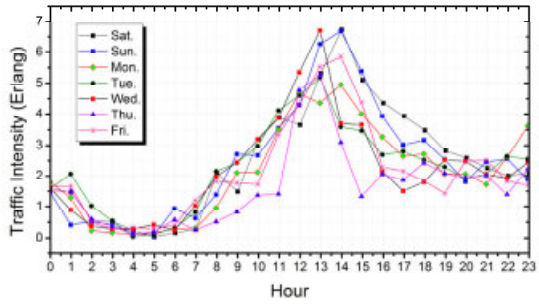


Fig. 1 Traffic Profile of Operator 2

The figures show that each traffic profile has a peak from 12:00PM to 2:00PM, wherein this period, applying green communication idea may not be possible due to the high network load. However, the low traffic period observed between 12:00AM to

08:00AM is encouraging for applying green communication.

The traffic profiles are used in our experimental work. The average daily traffic intensity (Au) is about (λ*H) 0.025 Erlang/User, where λ (the number of request per hour) =2 calls\hour, and H (the average call duration) = 45 seconds. The accepted Grade of Service (GoS) (blocking rate) is 2%.

4. Experimental Setup

Based on the BSs locations we got from two operators in Palestine, we constructed the network in OPNET as shown in figure 3. These BSs cover the Palestine Polytechnic University area. Through sectoring, each BS supports 3 cells. Distances between BSs are shown in the figure. The tarffic intensity profiles are used to emulate the network load. They are mapped to VoIP packets in OPNET. From the intensity profiles, we extracted a number of users during each hour. Users are randomly distributed across the coverage area of the BSs. Users are stationary. Transmitted power is used accroding to specifications from both operators (20Watt and 28Watt). The Hatta model for suburban areas is used for path loss. We focus on the low traffic period between 12:00am and 7:00am

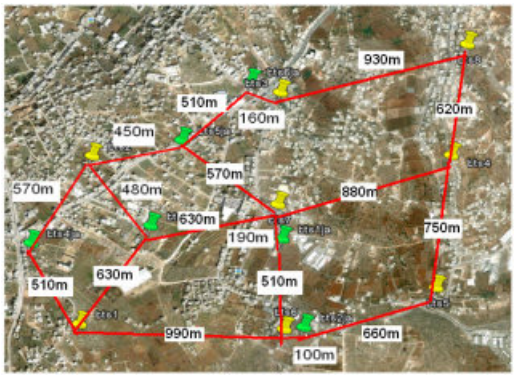


Fig. 3 BSs Distribution of both Operators

5. Simulation Results

5.1 Power Saving

Using equation 1, we computed the percentage of power that can be saved when green communication is applied during the low traffic period. Results are shown in Table 1.

TABLE X
PERCENTAGE OF POWER SAVED

Case	Percentage of Saved Power
One Network	**17.8%**
Two Networks (one Off – Roaming to one)	**33%**
Two Networks (Roaming to all)	**16.93%**

The results show that applying green communication by the two operators independently, each one can save about 17.8% of power. However, if the coverage area is serviced by only one of the two operators and the other switches off all his BSs, 33% of power can be saved. Finally, if some cells from both operators' networks

are switched off during low traffic period, 16.93% of power can be saved. Note that, in each of the three cases the SNR experienced by users will be affected due cells switching off and consequently the handoff among cells. The effect of green communication on the SNR experienced by cellular users is discussed in the next subsection.

5.2 SNR Analysis

With green communication, some cells will be switched off and their serviced users will handoff to other cells, which sometimes belong to different BSs. This is expected to influence the SNR user's experience. We addressed this issue for two cases. In the first case, one operator network is completely switched off and users handoff to cells of the second operator. The results are shown in figure 4.

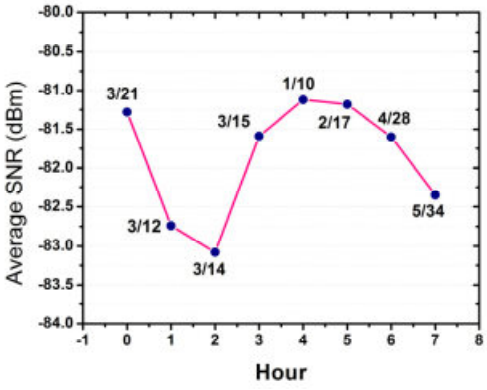


Fig. 4 Average SNR for the Case of One Network

The figure plots average SNR for all users during low traffic period. On the figure, we also show the fraction of users which experience SNR below -80dBm (a known value for acceptable QoS). The results show that, despite a 33% of power can be saved in this case (from section 5.1), some users may experience SNR below -83dBm when they handoff to the second operator network.

In the second case, some cells from both operators' networks are switched off. Figure 5 shows the average SNR experienced by all users with green communication. On the figure, we also show the fraction of users that experience SNR below -70 dBm. Interestingly, we have found that non of the users experience SNR below -80dBm. This is attributed to the fact that the consideration of cells of both networks, users have more potential closer cells to roam to. Noting the 16.93% power saving for this case, the results of figure 5 reveal the trade-off between power saving and QoS users will experience with green communication.

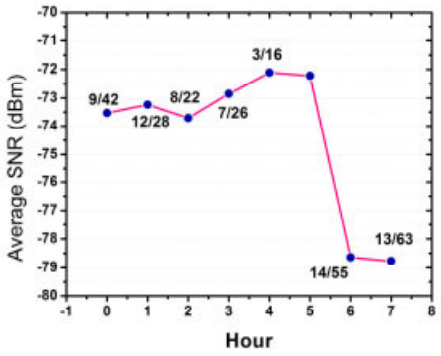


Fig. 5 Average SNR for the Case of Two Networks

6. Conclusion

In this paper, we investigate the possibility of applying green communication idea in Palestinian cellular networks. We considered the networks of two operators. The results show that a good amount of power can be saved with green communication. The results also show that in order to maintain a good SNR for all users, collaboration between operator's networks is necessary in order to handoff users to close cells when their original ones are required to switch off. Our next step is to work on development of policies and protocols to realize the green communication idea.

Acknowledgment

The author gratefully thanks both Palestinian cellular operators Jawwal and Wataniyya for making a lot of information available to us during the course of this work.

References

[1] M. A. Marsan and M. meo, *Energy Efficient Management of two Cellular Access Networks,* GreenMetrics2009, Seattle, WA, USA, June: 2009.

[2] Green Communications - Wireless@Virginia Tech

[3] M. A. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, *Energy-Aware UMTS Access Networks*,. In the Proc. of First International Workshop on Green Wireless, W-GREEN 2008, Lapland, Finland, Sept.: 2008.

[4] S. Vadgama, *Trend of green wireless access,* April:2009.

[5] M. A. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, *Optimal Energy Savings in Cellular Access Networks,* In the Proc. Of First International Workshop on Green Communications, Dresden, Germany, June: 2009.

[6] A.P. Jardosh, K. Papagiannaki, E.M. Belding, K.C. Almeroth, G. Iannaccone, B. Vinnakota, *Green WLANs: On-demand WLAN Infrastructures,* Journal of Mobile Networks and Applications, December: 2008.

[7] M.A. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, *A Simple Analytical Model for the Energy-Efficient Activation of Access Points in Dense WLANs,* In the Proc. of 1st International Conference on Energy-efficient Computing and Networking, e-Energy 2010, April: 2010.

Runtime Replica Consistency Mechanism For Cloud Data Storage

Mohammed Radi
Computer science department, Faculty of applied science
Alaqsa University Gaza
Moh_radi@alaqsa.edu.ps

Abstract

A cloud computing is becoming increasingly popular; cloud storage services attract more attentions for their high security and availability with a low cost. Cloud storage is expected to become the main force of the future storage market. As a key technology of cloud computing, replication faces new challenges, especially replica consistency. The intrinsic characteristic heterogeneous of cloud applications makes their consistency requirements different where the consistency requirement of certain application changes continuously at runtime. This paper presents a Runtime Replica Consistency Mechanism for cloud data storage to achieve a dynamic balance between consistency and performance. Evaluation result show that the propose mechanism guaranteeing the consistency and decrease the overhead.

Keywords: *replica consistency; cloud storage.*

1. Introduction

Cloud computing is becoming a very familiar word in industry and is receiving a large amount of attention from the research community.

Cloud storage is emerging as a powerful paradigm for sharing information across the Internet, which satisfies people's mobile data demand anywhere and anytime. Rather than relying on a few central large storage arrays, such a cloud storage system consolidates large number of geographically distributed computers into a single storage pool and provides large capacity, high performance storage service at low costs in unreliable and dynamic network environment [1].

Replication is one of the performance enhancing techniques for cloud storage system that has been widely used. Files are distributed across data nodes to achieve availability fault tolerance, scalability and performance. Unfortunately, the more replication increases the problem of inconsistent replicas. To solve the replica consistency problem, a replica consistency mechanism is needed. There are two traditional approaches that can generally be used as to how implement consistency management in large scale systems. The first approach is *lazy-copy based protocol (pull-based)* which transfers the updates from the original resource to the replicas only when accessing the

replica. This way can save a lot bandwidth resource because update transferring occurs only when accessing to the replicas but it cuts down the availability of replicas and increase replicas access time. The second approach is *aggressive-copy based (push -based)* in which insures all replicas to be updated immediately by transferring, the updates to all replica once the original replica is updated. This way can grantee the availability of up-to date data all the time but the maintain cost increase significantly. Lazy based and aggressive based are only suitable for particular scenes.[2]

In the cloud computing environment the customer of cloud storage is homogeneous. Some application need lazy consistency and some need aggressive consistency, and the consistency requirement is inconsistent and change at runtime. The mechanism for replica consistency should be suitable for different application and consider this consistency requirement changes at run time. [1,3,4,5]. This paper focus on introducing a Runtime Based Replica Consistency Mechanism (RBRC) for cloud storage to achieve a dynamic balance between consistency and performance.

The rest of the paper is organized as follows. In Section 2, we present the related works. Section 3 a runtime based replica consistency mechanism. the evaluation is presented in section 4 Finally , we conclude the paper in Section 5.

2. Related works

There are many consistency models proposed in the distributed systems and database literature. Common references in the DB literature include [6,7,8] Also In distributed systems, [9] is the standard textbook that describes alternative consistency models as well as their trade-offs in terms of consistency and availability. In data grid environment many replica consistency models have been proposed[2,10, 11, 12]. Our work extends these established models by allowing levels of consistency to be defined and adapting the consistency guarantees at runtime.

Strong consistency is expensive not just in the transaction cost, but also in terms of replicas' availablity and system's performance. Not all applications need strong consistency guarantees. However, eventual consistency may result in high penalty cost caused by false operations. Therefore, researchers pay attentions to the balance between consistency, availablity and performance.

Ximei Wang propose an application-based adaptive mechanism of replica consistency in cloud data storage they divide the consistency of applications into four categories according to their read frequencies and update frequencies, and then design corresponding consistency strategies. The results show that the mechanism decreases the amount of operations while guaranteeing the application's consistency requirement.[1]

Kraska proposes a strategy that system can switch the level of consistency between serializability and session consistency dynamically according to running condition in the cloud[4]. It divides the data into three categories, and treats each category differently depending on the consistency level provided. The consistency level will be changed accordingly while the data's impact changes continuously at runtime.

Islam proposes a tree-based consistency approach that reduces interdependency among replica servers by introducing partially consistent and fully consistent states of cloud databases. The tree is formed such a way that the maximum reliable path is ensured from the primary server to all replica servers [5].

Ruay-Shiuang proposes an adaptive replica consistency service for data grid[10]. The strategy treats replicas differently according to the access frequency during the initializing process. The original replica and first level replicas can be updated immediately. If access frequency exceeds a predefined threshold, the second level replicas are updated immediately, too. If not, the replica is only updated when it is accessed.

Dongmei Cao proposes an adaptive consistency model for grid according to access frequency[13]. Compared to [10], the most important improvement is allowing system to switch consistency level automatically at runtime.

Based on asynchronous aggressive update propagation technique, Radi, M propose a scalable replica consistency protocol to maintain replica consistency in data grid. In the propose protocol the high access weight replicas updated faster than the others.[12]

Ghalem compares pessimistic consistency with optimistic consistency, and combines these two existing approaches [2]. It divides replicas into several sites. Optimistic principals are used to ensure replica consistency within each site. Whereas, global consistency is covered by the application of algorithms inspired from the pessimistic approach. Some of the above researches don't allow the system to change consistency level automatically at runtime, so they can't achieve the dynamic balance between consistency, availability and performance. Some partition the consistency level continuously, so the switch transaction cost is high. And in some works, the metric is selected unilaterally, so it can't be the very representative for an application. In order to avoid the above problems, the adaptive consistency mechanism proposed in this paper is based on read frequency and update frequency. System can select a suitable strategy dynamically according to these two metrics at runtime.

3. Runtime consistency mechanism in cloud

3.1 Model structure

In this paper the management of replica adapts a single master nodes for each data item or file in which there exist single master copy which is the origin of the file and the other replicas are secondary replicas. Figure 1 show the overview of the system architecture.

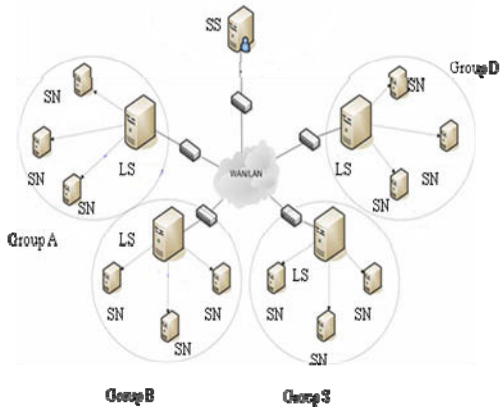


Figure 1: Over view of the system

We assume the SS, LS and SN are three main nodes of cloud sites that have more systems and storage resources. The super server (SS), where the original data are stored, can be modified by end users through data intensive applications. Several replica nodes located closely are organized into one group. Each group has one server consider as Local Server (LS) and other nodes in the group is consider as a Secondary Node (SN). LS is responsible for the consistency service within its group. A LS is responsible for executes a corresponding operation according to the replica consistency mechanism. The SN's are the replica holders, one of the secondary node in each group will act as a LS node when the previous one breaks down.

In our model both LS and SN can receive a read request from end user and only A SS can be modified by end users and if a LS or SN receive update request it forward it to the SS to process it.

3.2 Replica consistency architecture

In order to provide the required functionalities of the single master replication, Replica Consistency Service (RCS) architecture is proposed. Local Replica Consistency Service (LRCS) and Replica Consistency Catalogue (RCC) are the main components of the architecture:

- Local Replica Consistency Service (LRCS): it is responsible for updating its local replica and relay the update propagation process if necessary.
- Replica Consistency Catalogue (RCC) is used to store the metadata inclides information of all SN including the physical poison, update and read frequencies.; this metadata will used by the RCS.

The interactions between the above components are shown in Figure 2.

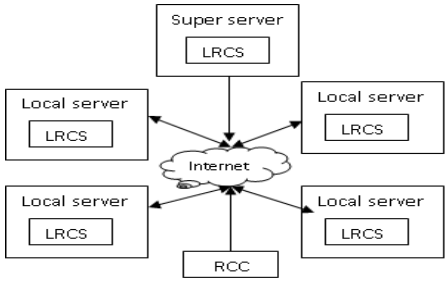


Figure 2: Replica Consistency Service Architecture

This interaction can be explained through a simple case user wishing to update a master file Fi. In a basic scenario, a user passes the update request and the target file to SS. The LRCS updates the local replica, and reflect a consistent view to its user, run the first step of the run time algorithm in which SS LRCS inquires the RCC about the replica LSs . After that SS send the update information to all LS. If the local server receives the update request it only passes this update request to the SS.details of the first step is shown in figure 3.

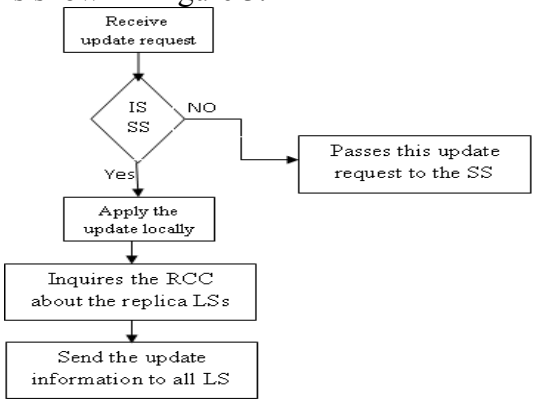


Figure 3: first step of the consistency algorithm

When the LS receive the update information, first it applies it at local replica and reflects the updates to its users, then it runs the second step of Runtime algorithm as show in figure 4. In the second step each local server inquires the RC about the read frequency of each replica in its group. Then it divides the replicas into two sup group depending on the read frequency. The replicas with high read frequency will be updated aggressively and the replicas with low frequency will be updated in lazy.

If the SS or the LS receive any read request from the user it directly allow the user to access the file, but if a SN receive a read request it first will check its read frequency, if the read frequency is high it directly allow the user to access the file but if the read frequency is low, then the LRCS pull the LRCS at LC for the last update, and it allow the user to access the replica only after it apply all the missing updates locally as shown in figure 5.

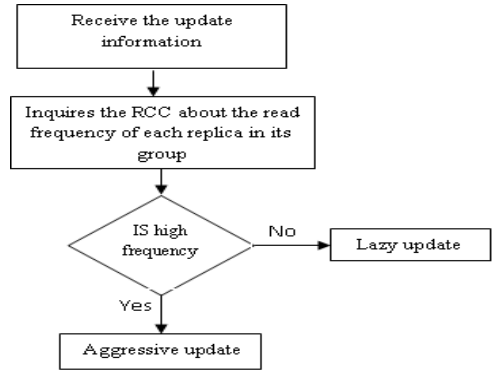


Figure 4: second step of the consistency algorithm

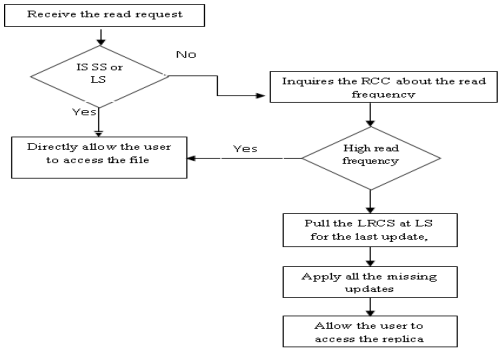


Figure 5: first step of the consistency algorithm

The runtime consistency mechanism divides the nodes into two categories according to the read frequency. The secondary nodes with high read frequency will follow the aggressive copy, and the low read frequency will follow the lazy- copy, while all master nodes will be updates in aggressive-copy. The ready frequency is classified as high if it exceeds a threshold value, and it classify as low if its less that a threshold value. Threshold value can be determined by cloud administrator.

4. Evaluation

We have implemented a runtime based replica consistency mechanism using OptorSim [14], a simulator for Data Grids. OptorSim was developed by the European Data Grid (EDG) project. It provides users with the Data Grids simulated architecture and programming interfaces to evaluate and validate their replication strategies. There are several critical omponents designed and implemented in OptorSim, including computing element (CE), storage element (SE), resource broker (RB), replica manager (RM), and replica optimister (RO), and so on. CEs and SEs are used to execute grid jobs and store files respectively.

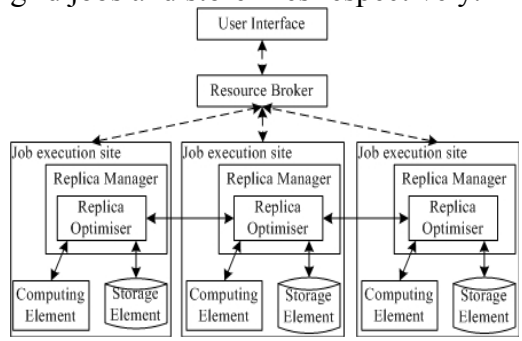


Figure 6. Basic architecture of OptorSim

In order to study the consistency we modify Optersim to satisfy our demand, and then compile our consistency mechanism on it. Each group is connected through the Internet. The Intra-region and inter-region network bandwidth are 1000Mb/sec and 500Mb/sec respectively.

Parameter	**Value**
Number of Jobs	500
Job Delay (ms)	25000
Max CE Queue size	200
File Processing Time (ms)	100000
Number of experiments	100
Each File Size (GBytes)	10
Number of Replica Modifications	100
Access Threshold	30

In order to study the runtime replica consistency mechanism we choose to compare our mechanism with lazy-

copy based protocol aggressive-copy based in term of average file access time , number of replication, percentage of requesting up-to-date date . File access time is defined as the real time duration that a CE spends for accessing one file including file replication time and file processing time. The number of replication is the number of replications needed to run the replica consistency mechanism. The higher number of replications means the more file transmissions may be taken place. It may consume a considerable amount of network bandwidth. percentage of requesting up-to-date date is defined percentage that the application accesses up-to-date data in time interval τ to be the representative of consistency requirement of an application, and the overall update amount to be the representative of transaction cost.

Recall that in the mechanism with higher number of replications file transmissions may be taken place. Also It may consume a considerable amount of network bandwidth. Figure 7 represents the number of replications for the three consistency mechanisms. For lazy protocol the number of replication is very small and the aggressive-copy based protocol the number of replication is very high and it may waste too much network resources on invalid replications because some replicas may never be accessed. Compared with Aggressive-copy based protocol, our mechanism could lower the number of replications without wasting valuable network bandwidth. And it take not so much number of replication than lazy- copy mechanism.

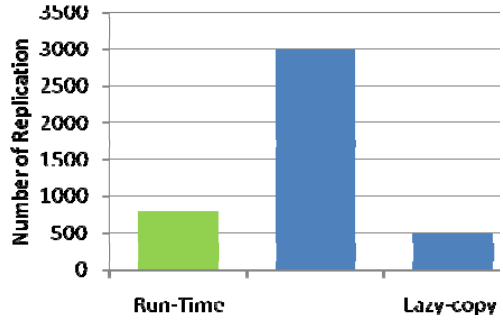


Figure 7: Number of replication

Figure 8 shows the comparison of the percentage that read a latest data every interval τ between lazy-copy, aggressive–copy and our consistency mechanism. Aggressive-copy almost guarantees that every access read the latest data. Lazy-copy mechanism guarantees weaker consistency, so the percentage is lower than strong consistency obviously. The run-time mechanism give a percentage in between the lazy-copy and aggressive – copy.

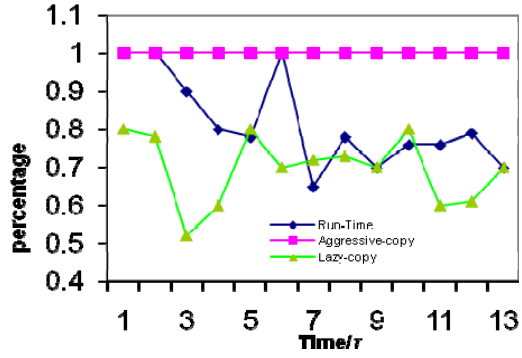


Figure 8: Percentage of accessing the up-to-date data

The average file access time have evaluated as shown in Figure 9. Compared with the Lazy-copy based protocol, run-time can reduce file access delay time significantly because of our shorter replication time. As for Aggressive-copy based protocol, it copies the up-to-date replica in its region all along, therefore the file access delay time is equal to the file processing time without suffering form the long replication delay time due to the consistency problems.

Compared to aggressive copy mechanism, the run-time consistency mechanism proposed decreases number of replication significantly while the needs of application for consistency are mainly satisfied. And our consistency mechanism guarantees higher percentage of read a latest data than lazy-copy and decrease the average file access time. Consequently, we get a better balance between consistency and performance.

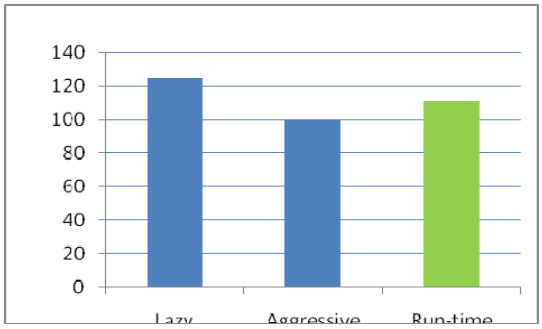


Figure 9: .Average file access time

5. Conclusion

This paper presents a Runtime Replica Consistency Mechanism for cloud data storage aiming achieve a dynamic balance between consistency and performance. The runtime consistency mechanism divides the nodes into two categories according to the read frequency. The mechanism maintain the replica consistency of some nodes in an aggressive way and some other node in a lazy way according to the read frequency. Evaluation result show that the propose mechanism guaranteeing the consistency and decrease the overhead

References

[1] Qingsong Wei, Bharadwaj Veeravalli, Bozhao Gong,Lingfang Zeng, Dan Feng, CDRM: A Cost-effective Dynamic Replication Management Scheme for Cloud Storage Cluster. 2010 IEEE International Conference on Cluster Computing, 188 – 196.

[2] Ghalem, B, S. Yahya. A Hybrid Approach for Consistency Management in Large Scale Systems. Proceedings of the International conference on Networking and Services, Page(s): 71

[3] Ximei Wang, Shoubao Yang, Shuling Wang, et, al, An Application-Based Adaptive Replica Consistency for Cloud Storage. 2010 Ninth International Conference on Grid and Cloud Computing.

[4] Kraska, T, M. Hentschel, et al. Consistency Rationing in the Cloud: Pay only when it matters. Proceedings of the VLDB Endowment, 2009, 2(1): 253-264.

[5] Islam, M.A.; Vrbsky, S.V.Tree-Based Consistency Approach for Cloud Databases Cloud Computing Technology and Science (CloudCom), 2010 IEEE

Second International Conference on Nov. 30 2010-Dec. 3 2010, 401 - 404

[6] P. Bernstein, V. Hadzilacos, and N. Goodman. Concurrency Control and Recovery in Database Systems. Addison Wesley, 1987

[**7**] M. T. Ozsu and P. Valduriez. Principles of Distributed Database Systems. Prentice Hall, 1999.

[**8**] G. Weikum and G. Vossen. Transactional Information Systems.

[9] A. Tanenbaum and M. van Steen. Distributed Systems: Principles and Paradigms. Prentice Hall, 2002.

[10] Ruay-Shiung Chang, Jih-Sheng Chang. Adaptable Replica Consistency Service for Data Grid. Third International Conference on Information Technology: New Generations (ITNG'06). 2006.

[11] Cao DongMei. The Research of Replica selection and consistency for Grid[D]. WuHan: Huazhong University of Science and Technology,
2007.

[12] Mohammed Radi ; Ali Mamat ; M. Mat Deris ; Hamidah Ibrahim ; Subramaniam Shamala, Access Weight Replica Consistency Protocol For Large Scale Data Grid, Journal of Computer ScienceYear: 2008 Volume: 4 - Issue: 2

[13] Cao DongMei. The Research of Replica selection and consistency for Grid[D]. WuHan: Huazhong University of Science and Technology, 2007.

[14] OptorSim: A Replica Optimiser Simulation.
http://edg-wp2.web.cern.ch/edg-wp2/optimization/optorsim. html.

Application of Computer Simulation for Optimizing Branchless Banking Opportunities via Cell Phones

Ashraf Al-Astal
Information Technology and Telecom Expert, Palestine
ashraf.astal@ieee.org

Abstract

Mobile Financial Services (MFS) are new phenomena in the world of Mobile Commerce (m-Commerce) which helps customers to interact with a bank via mobile device and makes banking virtually anywhere on a real-time basis a reality. This study investigates the impact of adopting MFS applications on minimizing service channels costs for Palestine Islamic Bank (PIB) in Khan Younis. Two types of models to analyse and evaluate the impact of adopting banking servicing opportunities via cell phones are presented. The first is a computer simulation model used for shedding some light on how inputs may affect the responses of interest. The second depends on the outputs of simulation experiment for finding the optimum combinations of input parameters by following Response Surface Methodology (RSM) assuming certain level of customers representing the early adopters will use MFS.

Keywords: *Arena Simulation, Computer Modelling, Optimization, Response Surface Methodology, Mobile Commerce, Mobile Financial Services, Operation Research.*

1. Introduction

Service industry has been developing rapidly and receiving more attention in the recent years by system modelers. Customer satisfaction is a growing concern in service industry settings such as banks, hospitals, and call centers. High variability in demand is prevalent in the service industry, and customers still expect to be served promptly when they arrive [1]. Therefore, there is a need for efficient staff utilization with minimal possible cost, taking into account varying demand levels for the day of the week, or even for the time of the day.

Improving customer satisfaction and service levels usually requires extra investments. To decide whether or not to invest, it is important to know the effect of the investment on the waiting time, and service cost. Usually managers and decision makers seek to balance between the service and waiting time cost to offer the best service with minimal cost [2]. Figure (1) shows the relation between these costs and how to obtain the minimum aggregate cost and optimal capacity.

Branch offices historically have been one of a bank's major costs as well as its main contact points with retail customers. This makes branches a logical target for efforts to cut costs and increase productivity. The movement to increase productivity and cut costs in branches led to the development of new channels for servicing opportunities, these opportunities offered the potential for reducing average transaction costs and less need for brick-and-mortar branch offices and tellers [3].

Mobile Financial Services (MFS) are a new phenomenon in the world of m-Commerce which helps customers to interact with a bank via a mobile device and makes banking virtually anywhere on a real-time basis a reality.

This study investigates the impact of adopting MFS applications on minimizing service channels costs and improving the performance of servicing levels for Palestine Islamic Bank (PIB) in Khan Younis.

Two types of models to analyze and evaluate the impact of adopting banking servicing opportunities via cell phones are presented. The first is a computer simulation model used for shedding some light on how inputs may affect the responses of interest. The second depends on the outputs of simulation experiment for finding the optimum combinations of input parameters by following Response Surface Methodology (RSM) assuming certain level of customers representing the early adopters will use MFS.

Arena Simulation Package is used to simulate the current banking system as to analyse the current performance and possible changes that could be made. By accurately simulating a process or system, the decision maker can see the outcomes of changes without implementing them in real-time, thus saving valuable time and resources. Design Experts 7.1 statistical package is used for constructing RSM plots and optimizing the input parameters for Tellers service channel.

2. System Description and Data Collection

PIB in Khan Younis is a small branch office of a national scale bank in Palestine. Services provided by this branch office include: savings, account deposits and withdrawals, transfer of funds, foreign currency exchange, and ATM services. The process of interest within this branch office is the service delivered to customers via tellers and ATM channels.

There are five counters in the tellers' area which can physically be opened

as shown in Figure (2), but current practice at the bank is to open two counters permanently -Permanent Tellers (PT)- with the additional counters being opened only during rush days and peak hours as needed. The additional counters served by Temporary Tellers (TT) called from the administrative employees.

The data collection stage gathers observations about system characteristics over time. The data used for the simulation model were collected during January, February, and March 2008. It consists of the total number of daily customer arrivals for the tellers and ATM service distribution channels during months of January and February 2008, the rates of customers arrival during day hours to the bank, average rates of service time for tellers and ATM resources, average waiting time for customers, average queuing lines length, cost for operating one teller and one ATM resource, average transaction cost for the transactions executed via tellers and ATM service channels. The data were collected from historical records, statistics, and supported with daily observations from the field.

3. Computer Simulation

The main goal of the simulation study is to examine new opportunities for distributing banking services via MFS. It will explore and evaluate the impact of MFS on improving customer satisfaction and reducing service costs for service channels.

The following assumptions were adopted for the purpose of modelling the system under investigation:

- Bank opens at 8:30AM and offers services for customers through tellers group until 1:00PM. At 1:00PM the bank closes but the customers in the bank at that time will be served. Customers come to the tellers' area with escorts.

- ATM services available for customers' usage between 8:00AM and 8:00PM only.

- Each teller serves customers from the same queue only and all tellers work at the same speed.

- During normal days, only two permanent tellers are working with the possibility to activate a temporary teller come from the back office if number of customers exceeds 30 customers. During rush days, two temporary tellers will be working in addition to the two permanent tellers but if the number of customers exceeds 50 customer in the lobby, a third temporary teller will be activated.

- Tellers do not have lunch hours, but do leave the counter from time-to-time. However, a teller will finish serving the current customers in the queue before leaving the counter.

- By default, all tellers can handle all types of customer transactions. There are no special service queues, but it might happen from time-to-time to let any teller serves a specific type of customers during day hours.

- There is no queue capacity limitation. This means that the building is large enough for the assumption of an infinite capacity is reasonable.

- When faced with several queues, customers tend to pick the shortest. However, all tellers will have an equal opportunity to serve any customer. All queues follow a First-In First-Out (FIFO) priority.

- Reneging and jockeying are neglected. However, a small percentage of customers may balk during normal and rush days.

- ATM machine may fail during period of operation due to power supply failure. If the failure was during branch office working hours, ATM customer will enter the bank and join tellers' area queues to be served; otherwise he/she may leave without being served.

Customers' arrival statistics shows that number of arrivals fluctuate throughout the rush day hours as shown in Table 1.

TABLE 1
OBSERVED AVERAGE ARRIVALS
DURING RUSH DAYS

Time Period	Observed Average Arrivals During Rush Days	
	Tellers Service Channel	ATM Service Channel
08:00-09:00	114	26
09:00-10:00	254	31
10:00-11:00	287	49
11:00-12:00	388	52
12:00-13:00	293	54
13:00-14:00	-	35
14:00-15:00	-	15
15:00-16:00	-	13
16:00-17:00	-	12
17:00-18:00	-	14
18:00-19:00	-	12
19:00-20:00	-	11

4. Experimentation and Results

The study consisted of two major phases. The first phase handled problem formulation and objectives of the overall project plan including collecting the data for the simulation experiment, simulation model

conceptualization and translation, in addition to verification and validation. The second phase handled experimental design and run for the simulation experiment. The outputs of the simulation experiment at this phase were used to RSM plots for the responses of interest. RSM used as an optimization technique for the main inputs of interest.

4.1 Experimental Design

A discrete event simulation model was created using Arena Simulation Package version 7. The model follows individual customers as they move through the bank teller/ATM systems. Flow of a customer would be as follows: Customer A arrives at the bank according to the arrival rate at a particular time of day. He enters the building and joins the shortest waiting line queue for service (It was assumed that 5% of the total customers may balk during rush days). Each time a teller becomes available, when they have finished serving a customer, the teller will call the next customer in line to approach the counter for service. When Customer A is called to a counter, he approaches and is serviced by the teller. Once service has ended, Customer A leaves the bank. The teller who serviced Customer A calls the next waiting customer to the counter. Number of operating tellers was changing during day hours. By default, four tellers will be working until the number of waiting customers in the lobby exceeds 50 customers, which is the point at which a fifth teller will start working until customers become below 30.

An essential step when designing any simulation model is to simulate the process over a specified period of time. It is important to tell Arena how long to run the simulation model and how many replications we need.

Most simulations can be classified as either terminating or steady-state (long run) [4]. A terminating simulation – in contrary to steady-state – for a single run will terminate according to some model specified rule or condition, which is the same as the case being analyzed in this study. The bank opens at 8:30AM with no customers present meanwhile ATM machine starts operating at 8:00AM. The bank closes its doors at 1:00PM, and then continues its operation until all customers are flushed out. The ATM machine continues its operation until 8:00 PM as it was not possible to collect data after this time limit; therefore the single simulation run was selected to start from 7:00AM to 9:00PM (840 minutes) with a one hour time margin at the beginning and at the end of each work day for the purpose of monitoring and validating customers' behaviour during the simulation run.

Each simulation run is replicated for 100 times, simulating a 100 day of operation for each combination of design points shown in table (2) that will be discussed in section 4.2. The service processing rate is assumed to

be homogeneous for all the tellers. Banking administrations usually circulate employees as to enhance their skills and improve performance levels; therefore it is assumed that all staff members have similar skills and performance levels. The service process is modelled using the triangular distribution as its parameters are fairly easy to understand [4]. Figure (3) illustrates minimum, most likely, and maximum service times for Tellers and ATM service channels during rush days.

4.2 Building RSM Plots

The use of Response Surface Methodology and Meta-Model procedure was involving the following steps: (1) Using 2^k Factorial designs for screening, (2) Using Central Composite Design (CCD) for building second order regression models, and (3) Optimizing model parameters during early adopters' stage.

The 2^k Factorial experiment design was developed with the physical experiments in mind (like industrial applications) and can easily be used in computer simulation experiments as well [5]. It is useful for shedding some light on which input parameters (factors) are most important and how they affect the responses of the experiment. The 2^k Factorial experiment design technique is based on identifying two values, or levels, of each of model input factors. There is no general prescription on how to set these levels, but it is important to set them to be "opposite" in nature but not so extreme that they are unrealistic [5].

If we have k input factors, then will have 2^k different combinations of the input factors, each defining a different opportunity of the model. Referring to the two levels of each factor as the "–1" and "+1" level, this can form what is called a design matrix describing exactly what each of the 2k different model opportunities are in terms of their input factor levels. Each row represents a particular combination of factor levels, and is called a design point.

For this study, set of opportunities were developed for modifying Permanent and Temporary Tellers (PT & TT) (factors A and B respectively) staffing levels under two distinct levels of MFS usage (factor C) as to examine the impact of different combinations on the responses of the model (like average total cost, average waiting time … etc.), thus we have three input controls (k = 3 factors), and 2^3= 8 runs representing cubic points for the full factorial experiment as shown in Fig. (4).

Permanent and Temporary Tellers staffing levels were considered to be

controlled input parameters (factors) for the simulation experiment as we can control the values of these factors meanwhile MFS usage factor was found to be *uncontrolled* as we do not have clear image on the percentage of usage at any specific point of time.

Central Composite Design (CCD) is a two level (2^k) factorial design, augmented by n_0 *Center Points*, and two *Star (Axial) Points* positioned at ($\pm\ \alpha$) for each factor as shown in Figure (5) for a three factors experiment design. Setting ($\alpha = 1$) locates the star points on the centers of the *faces* of the cube, giving a Face centered Central Composite (CCF) design.

CCD is useful in RSM for building a *second order* (quadratic) regression model for the response variable without needing to use a complete *three* level (3^k) factorial experiment for evaluating main and quadratic effects and interactions.

Table (2) shows the design matrix after using CCD for both normal and rush days configurations. Arena Process Analyzer was used to run the simulation experiment for these design points and record the responses for: (1) Tellers channel cost and wait time, and (2) ATM channel cost and wait time as a result for using MFS.

The outputs of Arena Process Analyzer were then processed by Design Experts (DX) software [6] in order to perform statistical analysis based on CCD experiment design as to construct response surface plots in addition to regression models for the responses that might be used later by decision makers to predict the outputs of simulation experiment. The use of meta-models can help to find the combination of input factor values that *optimizes* (i.e., minimizes or maximizes, as appropriate under some constraints) responses which is the task that will be handled in the next step.

4.3 The Results

The responses of the basic simulation model were compared to the MFS-based simulation model responses. The results shown in table (3) indicates that the 36.5% of MFS early adopters will reduce customers waiting time in the tellers area to (10.82 minutes) on average with a total cost of (487.30 $/Day), which is much less than the total cost of tellers area without using MFS (1735.47 $/Day). In addition, ATM service delivery channel cost might reach (76.08 $/Day) with an average service time of (3.52 minutes) which is again much less than current basic configuration. At this level of MFS usage, there is a possibility to

generate (38.01 $/Day) net profit from the current customers base during rush days.

TABLE 2
OBSERVED AVERAGE ARRIVALS
DURING RUSH DAYS

Design Point	Design Point Type	Factor 1: Number of Permanent Tellers		Factor 2: Number of Temporary Tellers		Factor 3: Percent of Usage for MFS	
		Coded	Natural	Coded	Natural	Coded	Natural
1	Factorial	-1	3	-1	0	-1	20
2	Factorial	+1	5	-1	0	-1	20
3	Factorial	-1	3	+1	2	-1	20
4	Factorial	+1	5	+1	2	-1	20
5	Factorial	-1	3	-1	0	+1	60
6	Factorial	+1	5	-1	0	+1	60
7	Factorial	-1	3	+1	2	+1	60
8	Factorial	+1	5	+1	2	+1	60
9	Center	0	4	0	1	0	40
10	Axial	$-\alpha^{(*)}$	3	0	1	0	40
11	Axial	$+\alpha^{(*)}$	5	0	1	0	40
12	Axial	0	4	$-\alpha^{(*)}$	0	0	40
13	Axial	0	4	$+\alpha^{(*)}$	2	0	40
14	Axial	0	4	0	1	$-\alpha^{(*)}$	20
15	Axial	0	4	0	1	$+\alpha^{(*)}$	60

(*) $\alpha = 1$

TABLE 3
COMPARING BASIC AND MFS_BASED
FOR RUSH DAYS

Simulation Model	PT (No.)	TT (No.)	MFS Usage (%)	Teller Area Total Cost ($/Day)	Teller Customer Wait Time (Minutes)	ATM Area Total Cost ($/Day)	ATM Customer Wait Time (Minutes)
Basic	4	1	0%	1735.47	43.03	486.52	26.33
MFS	4	1	36.5%	487.30	10.82	76.03	3.52

DX software was used for locating the sweet spot area with multiple responses that satisfies the constraints listed in table (4) and help to find the combination of input factor values that optimizes responses in order to achieve the required goals.

TABLE 4
CONSTRAINTS AND GOALS TO BE
ACHIEVED FOR A RUSH DAY

Name	Goal	Lower Limit	Upper Limit
PT (Number)	is in range	3	5
TT (Number)	is in range	0	2
MFS (%)	is equal to 36.50	20	60
Teller Area Total Cost ($/Day)	minimize	214.26	2562.28
Customer Wait Time – Tellers Area (Minutes)	is target = 10	10	15
ATM Area Total Cost ($/Day)	minimize	17.78	249.24
Customer Wait Time – ATM Area (Minutes)	is in range	0.29	13.23
Net Profit ($/Day)	Maximize	20.59	62.42

The output of DX shows that the stationary point of the fitted response surfaces X_o = (PT, TT, MFS) = (3.99, 0, 36.5) which yields predicted mean responses of Y_o = (Y1, Y2, Y3, Y4, Y5) = (449.43, 9.99, 65.04, 3.17, 37.10) a maximum in the experimental region as shown in table (5) with a solution desirability of 86.865%.

TABLE 5
THE DX OPTIMIZATION PROCESS
OUTPUT FOR A RUSH DAY
CONSRAINTS AND GOALS

Solution Number	X1: PT (No.)	X2: TT (No.)	X3: MFS (%)	Y1: Tellers Area Total Cost ($/Day)	Y2: Tellers Area Customer Wait time (Minutes)	Y3: ATM Area Total Cost ($/Day)	Y4: ATM Area Customer Wait Time (Minutes)	Solution Desirability (%)
1	***3.99***	***0***	***36.5***	***449.43***	***9.99***	***65.04***	***3.17***	***86.865***
2	3.99	0	36.5	450.62	9.99	65.04	3.17	86.861
3	3.99	0.12	36.5	452.21	10.00	65.04	3.17	86.857
4	3.99	0.13	36.5	452.41	9.99	65.04	3.17	86.851
5	3.99	0.25	36.5	455.03	10.00	65.04	3.17	86.843
6	3.99	0.57	36.5	462.41	10.00	65.04	3.17	86.814
7	3.99	1.55	36.5	484.96	9.99	65.04	3.17	86.725
8	3.99	1.74	36.5	489.48	10.00	65.04	3.17	86.701
9	3.99	1.82	36.5	491.34	9.99	65.04	3.17	86.690

If we take the combination of nearest integer values of PT and TT as X = (PT, TT, MFS) = (4, 0, 36.5), the output responses will be Y = (Y1, Y2, Y3, Y4, Y5) = (447.70, 9.93, 65.04, 3.17, 37.10) which is very close to Y_o as X is very close to X_o and lays within the solution desirability region as shown in Fig. (6) which leads us to conclude that X is the optimal solution for a rush day configuration.

Figures (7), (8), (9), and (10), shows the response surface plots of the tellers area total cost and customers waiting time, ATM area total cost and customers waiting time responses respectively at the level of 36.5% of customers are using MFS. These plots confirms the results of DX optimization output for the optimal solution as the lowest cost of tellers area seems to be achieved at 4 permanent tellers only with a waiting time between 5 and 15 minutes.

Figures (9), and (10)) explain that ATM channel total cost, and ATM channel customers waiting time responses is constant at all points since these responses are a function of MFS usage factor only.

5. Conclusion

Banking will no longer be constrained to conventional service channels. Mobile Financial Services (MFS) are new phenomena in the world of M-Commerce which helps customers to interact with a bank via a mobile device and makes banking virtually anywhere on a real-time basis a reality. MFS can be divided into two sub-categories: Mobile Payment (M-Payment) and Mobile Banking (M-Banking). The advantage for the customers lies in the fact that they does not need to carry cash. Therefore, MFS

can be seen as a promising M-Commerce application.

If implemented proficiently, MFS can help financial institutions and banks in Palestine to improve customer acquisition and customer retention, reduce total service costs for costly branch offices by migrating simple transactions away from these branches.

This study has shown how MFS may affect PIB operations in Khan Younis branch during normal and rush days. The simulation has been valuable in providing a flexible environment in which to model PIB for Khan Younis branch and gather information about the arrival patterns and key performance indicators which were necessary for running the simulation experiment and applying RSM as an optimization technique. The RSM is used to find the optimum levels of the considered factors to ensure a well-designed physical system.

By simulating the behaviour of the queuing systems in the bank for both types of days at the level of 36.5% of customers are willing to use MFS, the study concluded that:

- During Rush Days: It has be shown that providing 4 permanent tellers with a solution desirability of 86.86% will lead to the required minimum average of customers waiting time (9.93 Minutes) as to serve customers within 10 to 15 minutes at a total average cost for tellers service channel of (447.70 $/Day), which is much less than the total average cost for tellers area without using MFS (1735.47 $/Day), in addition for eliminating the need for a temporary tellers during day hours to be ready. At this level of MFS usage, the total average cost and waiting time for ATM service channel will be reduced to (65.04 $/Day) and (3.17 Minutes) respectively which is again much less than the average cost and wait time for ATM area without using MFS. In addition, this level of MFS usage will allow for an opportunity of a maximum of (37.10 $/Day) net profit that might be generated from the current customers base.

Good utilization of MFS will help to find new methods for payments and money withdrawals as to overcome the difficulties appeared in Gaza Strip area resulting from the limitations for supplying Israeli Shekels, US Dollars and Jordanian Dinars to the Palestinian banks which were used for day-to-day transactions and caused many difficulties for both customers and banks. It is important for Palestinian banks to learn from successful stories were applied around the world.

In conclusion, MFS is offering several benefits and added value to banks and customers alike. Mobile operators value M-Commerce applications much. Moreover customers' perception of MFS as a new innovation will affect the rate of adoption. Many innovations require a lengthy period of many years from the time when they become available to the time when they are widely adopted. Therefore, a common problem for many banks is how to speed up the rate of diffusion of MFS. The challenge for banking sector in Palestine is not to get unbanked to the bank, but to get the bank to the unbanked.

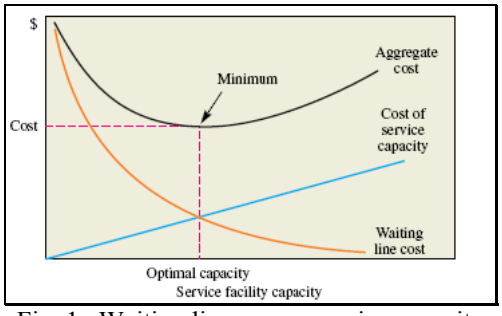


Fig. 1 Waiting line versus service capacity level trade-off
Source: (Chase, 2007)

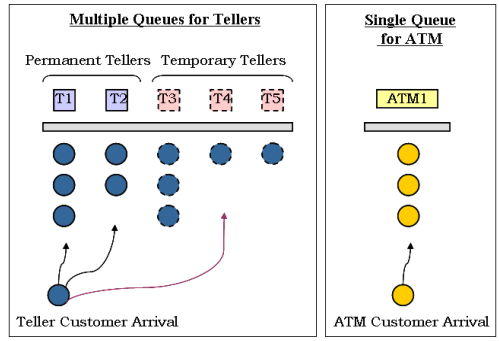


Fig. 2 Tellers and ATM services subsystems

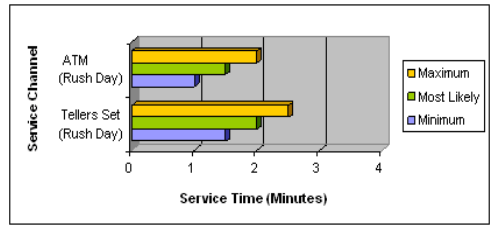


Fig. 3 Service rates for service channels

.

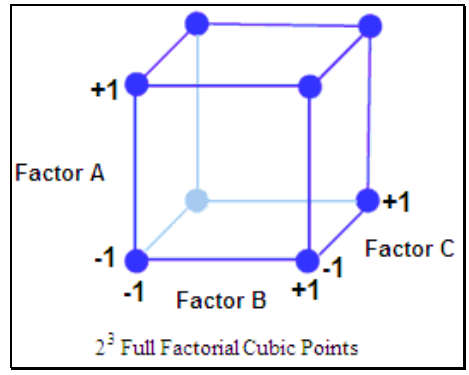


Fig. 4 Cubic points of 23 full factorial design
Source: Modified from (Sanchez, 2007)

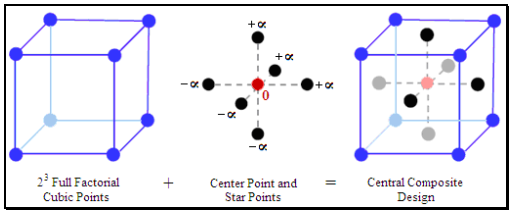


Fig 5 Construction of Central Composite Design (CCD)
Source: Modified from (Sanchez, 2007)

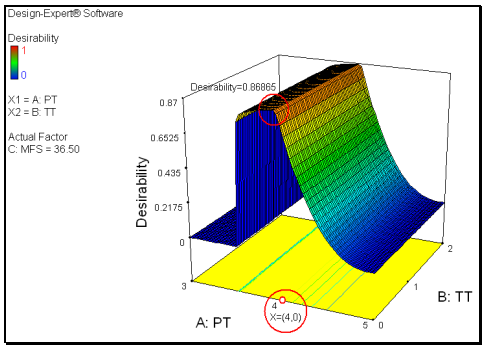


Fig 6 Desirability plot of rush day solutions

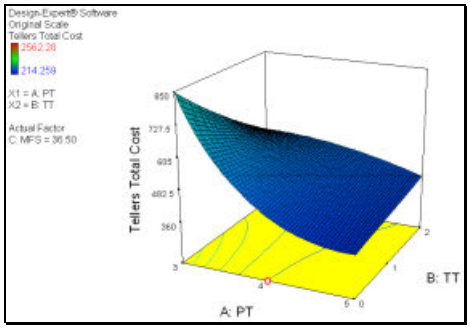


Fig. 7 Total cost response of tellers channel during rush day

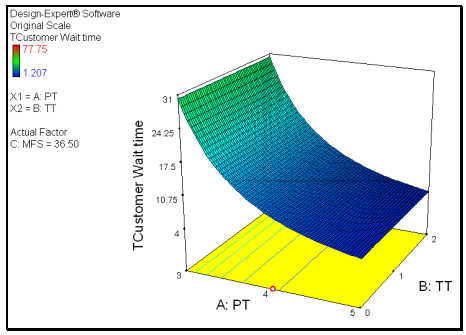


Fig. 8 Customers waiting time response of tellers channel during rush day

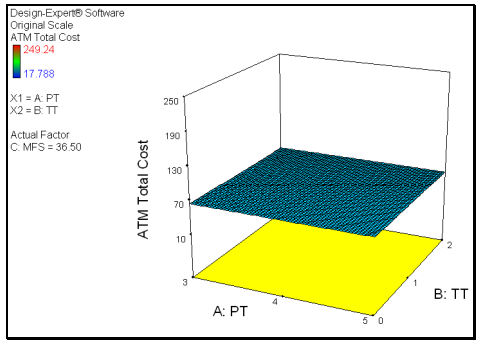


Fig. 9 Total cost response of ATM channel during rush day

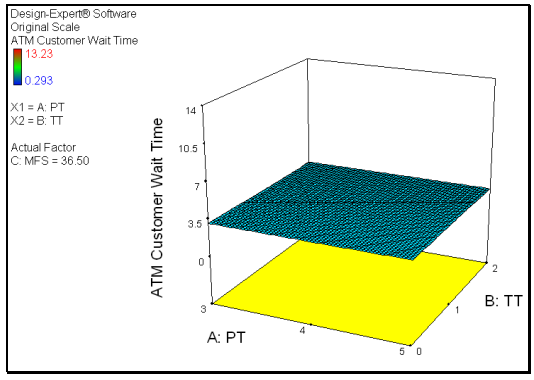


Fig 10 Customers waiting time response of ATM channel during rush day

Acknowledgment

I would also like to express my deep gratitude to Mr. Aziz Hammad, who works at the Palestine Islamic Bank (PIB) on which this study was based, for taking the time to support and provide me with the information I needed to complete this project. Additional thanks are due to Mr. Ahmad Fares, manager of PIB Khan Younis branch, also I would like to thank employees in the PIB, who provided me with data and answered my questions concerning the data collection.

References

[1] Chandra, W. and Conner, W. (2006), 'Determining Bank Teller Scheduling Using Simulation with Changing Arrival Rates', Research project, College of engineering, Pennsylvania State University, USA, retrieved on 13 November 2007 from website:

http://www.personal.psu.edu/wxc202/cv/Determining%20Bank%20Teller%20Scheduling_Wenny%20Chandra_Whitney%20Conner.pdf

[2] Chase, R. (2007), 'Operations Management', Second Edition, Publisher: McGraw Hill Inc.

[3] BCC. (2002), 'Global Markets for Retail Banking Technology: Retail Banking Solutions Overview (Increasing Productivity/Reducing the Cost of Services: New or Streamlined Business Process)', Business communications company, retrieved through subscription on 4 November 2007 from website:

http://bcc.ecnext.com/free-scripts/document_view_v3.pl?item_id=0279-106867&format_id=HTML

[4] Kelton, D., Sadowski, R. and Sturrock, D. (2004), 'Simulation with Arena', Third Edition, Publisher: McGraw Hill Inc., New York.

[5] Kelton, D. and Barton, R. (2003), 'Experimental Design for Simulation', Proceedings of the 2003 Winter Simulation Conference, USA.

[6] Statease. (n.d.), 'Design Experts V7.1 Software', Stat-Ease Incorporation, USA, Free full functional trial version for 45 days downloaded on 23 June 2008 from website:

http://www.statease.com/dx7trial.html

[7] Chase, R. (2007), 'Operations Management', Second Edition, Publisher: McGraw Hill Inc.

[8] Sanchez, S. (2007), 'Work Smarter. Not Harder: Guidelines for Designing Simulation Experiments', Proceedings of the 2007 Winter Simulation Conference, USA.

Mobile Learning Applications

Ramy I. R. Ashour
Al-Quds Open University – Palestine
rashour@qou.edu

Abstract

Last decade, mobile technologies have grown from a minor research to significant projects prevailed most of various lifestyles. Beside m-commerce, m-learning was one of the most interesting projects. Each project has illustrated how this technology can offer a new opportunity for learning that extends within and beyond the traditional teacher-led classroom. In fact, many higher education institutes have developed their own applications or adopted some commercial versions, yet they are successful only when developers understand the strengths and weaknesses of the technology beforehand, and integrate technology into appropriate pedagogical practices. This paper aimed to discuss the conceptual frameworks and the prerequisites of designing m-learning applications and resources giving a view of how to design a useful mobile application with limited-capabilities technology in education field. Experienced examples of a good practice in m-learning application are presented also in this paper.

Introduction

Today, the world is witnessing a significant growth of informatics technologies. People have to access information; over the past, they have had to obtain information from scientists, clergy, libraries and universities. From commerce, business, industry and medicals to every aspects in our life, and the education is one of the most important of those, whereas the conventional learning hasn't been able to fully adapt the growth (Ferioli & Van der Zwaan, 2009) with consideration about printed study resources and written assignments submitted manually to tutors who provide feedback. However, recently, students can access information and do their studying activities through the on-line learning (e-learning) environment without the need for such efforts (Gregson & Jordaan, 2009).

The need remains, because of the nature of PC, and internet has restricted the ubiquitous potential of e-learning to those moments when a learner is at home or at work in front of their PC. On the move, the learner cannot access the learning resources nor complete their course work (Motiwalla, 2007). Nowadays, the advancement of mobile devices and technologies presented during 2009

and 2010 especially with the introduction of new iPhone 3Gs and 4Gs, then iPad tablet, the mobile learning (m-learning) opportunities have increased highly (Fetaji & Fetaji, 2011); it is seen as an evolution of e-learning emerging the mobile device as a single integrated point of communication, and a useful access to information, applications and users (students/teachers) (Boja et al., 2009) (see figure 1).

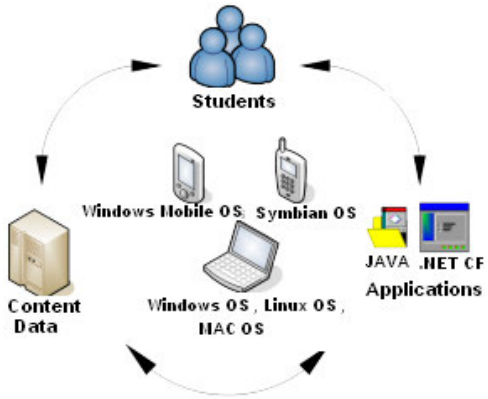


Figure 1: General architecture of m-learnin (Boja et al., 2009)

Harris, (2001) stated: "m-learning is the point at which mobile computing and e-learning intersect to produce an anytime, anywhere learning experience" (Pieri & Diamantini, 2009). In other words, m-learning overcomes the limitations of e-learning and meets the needs; it allows learners to access information and complete other course work even when they are away from their hard-wired internet connection.

M-learning theories and conceptual frameworks

Interest in mobile learning is growing in higher education as signified by the number of projects, conferences, scholarly journals, technical reports and books (Crow, Santos, LeBaron, McFadden, & Osborne, 2010). Many reviewed researched studies (Kennedy, 2003; Kukulska-Hulme & Pettit, 2009; Yordanova, 2007) have given encouraging results for using mobile technologies to support educational users in the teaching and learning process (Fetaji & Fetaji, 2011).

On the other hand, the communities cohering around mobile learning may still feel the need for standard and obvious theory of mobile learning as well as a definition (Traxler, 2009). Furthermore, theoretical justification is arguably even more important, when there is inadequate empirical evidence of effective learning with mobile technologies, guidelines for use should be theory-informed (Herrington & Herrington, 2007). Fishman, Soloway, Krajcik, Marx, & Blumenfeld, (2001) contended to set

theoretically grounded guidelines represent "a major impediment to the successful use of new technologies". Many research studies and projects have examined mobile learning from an identified theoretical perspective (Herrington & Herrington, 2007), Herrington and Herrington, (2007) introduced theories that are useful for guiding the design of technology-supported learning environments for higher order learning and were as a ground of foundation for some studies (J. Herrington, Herrington, Mantei, Olney, & Ferry, 2009).

- Behaviourist theory: Activities that promote learning as a change in observable actions.
- Constructivist theory: Activities in which learners actively construct new ideas or concepts based on previous and current knowledge.
- Situated learning: Activities that promote learning within an authentic context and culture.
- Collaborative learning: Activities that promote learning through social interaction.
- Informal & lifelong learning: Activities that promote learning outside a dedicated learning environment and formal curriculum.
- Learning and teaching support: Activities that assist in the coordination of learners and resources for learning activities.

Recently, Siemens, (2005) came out with a theory called Connectivism, it has been described as 'a learning theory for the digital age', and its characteristics include:

- Learning and knowledge rests in diversity of opinions.
- Learning may reside in non-human appliances.
- Capacity to know more is more critical than what is currently known
- Nurturing and maintaining connections is needed to facilitate continual learning.
- Currency (accurate, up-to-date knowledge) is the intent of all connectivist learning activities.

Nevertheless, in their conclusion, Cobcroft, Towers, Smith, & Bruns, (2006) indicated that also there was little attention being paid to developing specific frameworks to support the design of mobile learning. An initial attempt offered by (Sharples, Taylor, & Vavoula, 2005) suggested that a theory of mobile learning should be measured under the following criteria:

- Is it significantly different from current theories of classroom, workplace or lifelong learning?
- Does it account for the mobility of learners?
- Does it cover both formal and informal learning?

- Does it theorize learning as a constructive and social process?
- Does it analyze learning as a personal and situated activity mediated by technology?

M-Learning Application Issues

Mobile devices can be more easily integrated across the curriculum than desktop computers (Yordanova, 2007) and in a classroom environment without any extended requirements because of the environment infrastructures and the context of use. But, the success of m-learning is also limited to the hardware and software constraints of mobile devices. The lack of data input capability, limitation of processor speed, performance and memory storage, compatibility issues, limitation of file types supported, screen size and battery life (Maniar, Bennett, Hand, & Allan, 2008), directly influence the usability of mobile devices in the learning process.

However, technological capacity of all mobile devices has increased dramatically in the last few years. Nowadays, Screens are bigger and better, systems have more memory, and have more multimedia capabilities; as well as there are more convenient methods for data input (Chiu, Hung, & Street, 2009). Moreover, continuous advancing in mobile hardware technology, communication, the evolution of functionalities, ubiquitous availability of wireless networks and mobile devices are getting increasingly more powerful in terms of computing power and memory storage. Ongoing development of broadband wireless networks and the quick increase of power and capacity of cellular phones have enhanced the potential of mobile technologies in education (Boggs, 2002).

On the other hand, despite these advances of mobile technology, some obstacles exist which is still limited (Fetaji & Fetaji, 2011):

- Small screen size and low screen resolution.
- Low storage capacity and network bandwidth.
- Limited processor performance.
- Short battery life.
- Compatibility issues.
- Lack of data input capability.
- High - cost browsing through GPRS and 3G and 4G technologies.

Consequently, these limitations have shown some usability problems. The mobile screen is not equal to desktop screen. It has no sufficient space to display greater amount of information; the information may not

appear properly. Therefore, a vast amount of information in a small screen might affect the users' recognition. Small screen also restricts displaying lot of graphics. Due to the low graphic resolution and limited number of colors, the interface objects and multimedia information may appear despoiled and not obvious.

With the limited display quality, users need to focus on the environment rather than the application, so output is limited (Cavus & Ibrahim, 2009). The mobile application interface shouldn't become a scaled desktop application interface. This degraded in visual appearance of interface elements in mobile screens will negatively affect the quality and efficiency of user acceptance and understandability of the learning resources.

In other side, desktop applications cannot be accessed via mobile devices and be displayed same in a mobile screen. "What works well on a large screen does not necessarily work well on a small screen" (Kukulsha-Hulme & Traxler, 2007). Most existing computer based learning management systems still do not have access support for mobile devices, and there are deficiencies in cross-platform solutions of LMS (Fetaji & Fetaji, 2011). Moreover, many mobile browsers do not support scripting or plug-ins, and do not have available memory to display desktop pages and graphics. This directly influences the usability of mobile learning systems. Web content that is mostly the format of electronic learning content is poorly suited for mobile devices (Cavus & Ibrahim, 2009). So, this limitation will decrease the ability to display information in various multimedia formats.

These usability issues of mobile devices and learning must be considered and carefully examined during the usability testing of a mobile application in order to select an appropriate research methodology and reduce the effect of contextual factors in the outcomes of usability testing (Kukulska-Hulme & Pettit, 2009).

For the evaluation of m-learning activity, in their commercial study, Gregson & Jordaan, (2009) put some important questions introduced to the learners:

- What kind of learning objectives, and pedagogical approach is the activity suitable for?

- Are there specific technical pre-requisites that need to be met in order to make use of the activity, e.g. file types, network services?
- Were any relevant constraints identified when testing the activity in the community?
- How is the activity best delivered to the learner?
- How was the activity designed, and what resources were required?
- How was the activity evaluated?
- What were the learner and tutor reactions to this activity when it was tested?

Prerequisite concepts of designing mobile learning application and resources

Mobile devices such as mobile phones, PDAs and iPods can have more processing power, slicker displays, and more interesting applications than were commonly available on desktop machines one decade ago, and educators are quickly realizing their potential to be used as powerful learning tools.

However, to provide learners and teachers with better opportunities and enhanced learning outcomes, it is important to consider mobile issues discussed above before implementing application and designing the learning environment and resources (Park, 2011). This section (adapted from (Dochev & Hristov, 2006; Low & O'Connell, 2006)) suggested prerequisite concepts of good practice in mobile learning that should be a guidance for application developers and learning resources designer with a strong pedagogical basis depending on characteristics of mobile technology.

1. Compatibility and developing environment features

Mobile operating systems offer fewer application programming interfaces (APIs) than PCs do; developers need to be aware that not all PC functionality is supported in mobile systems. The various OSs available each have different advantages, but in all cases the functionality of mobile systems is more limited than that of PCs. Moreover, designers have to be educated and aware of various platforms; learning resources must be deployed as a wide range of devices as possible in same quality even that delivered on non-mobile platforms and test the resources across platforms. In this case, designers have to apply the most appropriate standards for usable, accessible and exchangeable format.

2. Performance and device resources

Limited system resources with narrow bandwidth have to be taken in consideration. Learning resources and contents have to be designed and provided as quickly with few processing as it possible. Developers have to determine where the processing will take place either in the local device or in the network server. Applications limited by processing rather than by bandwidth, is clearly better to be performed on the server, where decoding files such as (MP3) is more appropriate for local device processing. If the constraint is in the bandwidth, the goal will be to reduce the amount of data that has to be transmitted.

For processing, in real time processing, if the application has a lot of dada, developer has to reduce the need of processing during data transfer, such as compression. On the opposite side, if the application is computation-intensive but doesn't have much data, the goal will be to offload as much processing to the network as possible.

3. Memory available

A very important issue for the developer is working with limited size memory, where it is shared between stored data and active processing. Developer has to optimize the application software by removing unnecessary features, minimize program and reduce using recursive functions until it is absolutely necessary, this function would maximize memory stack. Moreover, to prevent memory leaks, the application should eliminate unnecessary memory allocation and free all allocations on exiting processes. Resources designers have to split the content into smaller objects or resources so that users can choose to store and/or download just the bits they need.

4. Small screen interface design

With up to 240 X 320 pixel and 3.75 inches (Maniar, Bennett, Hand, & Allan, 2008), the small and limited display size and resolution of these devices and interaction styles impose new interface designs. In this context, the interface has many constraints, needs to be simpler and might contain less number of components and objects. Developers must take care to eliminate unnecessary data from the screen. Often the appeal of an application in a PC lies in taking advantage of the display capabilities

and system graphics. In a handheld system, with its small, low-resolution screen and simple graphics, the application will have to be more limited in its video output. Here the challenge for the software developer is to take less and make the most of it to create a satisfactory interface for the user; Developer has to carefully layout pages and prevent scrolling into either dimensions, he has to chose more appropriate fonts to maximize readability and create graphics that are easy to view.

5. Saving power (battery life)

In mobile application, the power consumption is one of overriding issues, so developers should be aware of and use as low-power system features as they can. Mobile OSs typically provides power management features that allow for the partial shutdown when the system is in idle cycles. Therefore, it is important for the application to return control to the OS when it is waiting for a system resource. For instance, if the application needs input from a button on the keyboard, it should send an event, and then wait for system responce to inform it that the event has occurred. Doing so eliminates so-called “busy waiting” when the application does not return control to the OS while it is idling, that will save the power and enable longer use of the system between battery charges.

Examples of m-learning application

Mobile Interactive Learning Objects (MILO)

In their study, Holzinger, Nischelwitzer, & Meisenberger, (2005) presented a practical approach to m-learning for medical staff and students and call it "Mobile Interactive Learning Objects (MILOs)" which is used within a Mobile Learning Engine (MLE) that runs on mobile phones. MILOs can offer manifold possibilities for new kinds of communication and explorative learning. MILOs was structured the same way as Learning Objects for e-learning with taking into account some issues regarding to the limitation of the screen size and memory size. Different kinds of media were applied: figures and pictures, videos, audio and the most important, the possibility of the output of spoken text. Nevertheless, a MILO doesn't contain as much information as a traditional learning object (LO). Therefore, it splits up a

topic into separate MILOs, which are related but independent of each other. That will enhance video and audio streaming as well as in playback (see figure 2).

Moreover, through m-Learning, MILO can be primarily used during idle periods that may end abruptly. For example: a medical student, waiting for a bus, can decide to use this spare time for learning on a mobile phone until arrival at the training hospital.

Figure 2: Example screenshots for a MILO: Media-bar for the playback of the video

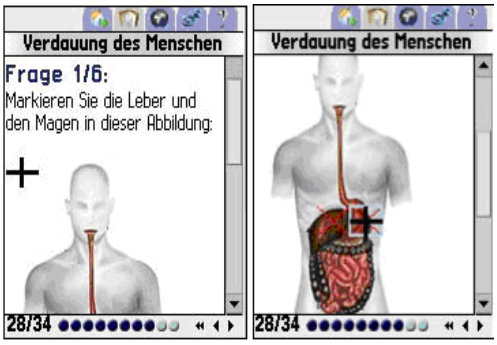


Figure 3: Example screenshots for a MILO: Interactive questions

Within MILOs, the system can define interactive questions (see figure 3). The answer to these questions will be analyzed and corrected by the application itself. The system can also define hints for question that if the users' solution to the questions are incorrect, the application provides a hint, which assists them to rethink about the problem and help them to find the correct solution. Punishment in the form of a WRONG message is replaced by encouragement and assistance. By solving the question, independently, the users get a feeling of success and increase their knowledge.

Mobile moodle (Moodbile)

Moodle is an open source Course Management System (CMS). It is also known as a Learning Management System (LMS) or a Virtual Learning Environment (VLE). It has become very popular among educators around the world. Moodle can provide a unique opportunity for students to enroll in social negotiation and mediation in the form of synchronous and asynchronous communication technology. Online communications are allowed for social negotiation and

mediation to occur across both time and distance (Wood, 2010).

Nevertheless, what happens if a student wants to read the forum posts while he/she is on the underground without wireless access?. Does the student need to pay to read a document in the virtual campus every time she/he wants to read it?, And what if she/he has a wifi access in the cell phone and wants to get all the data while she /he has free connection and review these data while she/he is on the go?

The point is that the students might want to access the data from the LMS when they are offline, and this is not possible in a web based scenario, too. One possible way to overcome this problem would be the use of web caching tools or RSS feed readers. But the data in the LMS is password protected, and many issues can appear even if we do not consider the security problems.

For the previous reasons, some institutes have utilized the open source of Moodle and create their own mobile extension for a LMS system (Forment, Guerrero, & S‹ínchez, 2008) and called it as (Moodbile) , where Moodle organization has designed a standard version of mobile Moodle. Figure 4 shows snapshots of different moodbile designs.

Figure 4: Screenshots of Moodbile

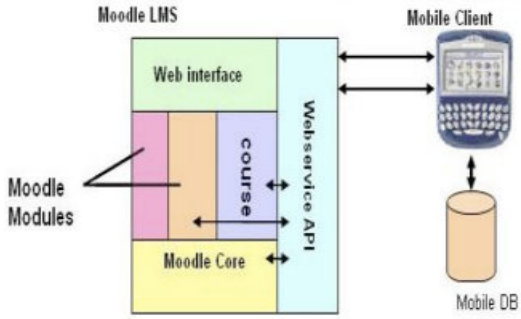


Figure 5: General system architecture of Moodbile (Forment, Guerrero, & S‹ínchez, 2008)

The design of a mobile extension for an LMS system takes two software design, server and client sides. The general system architecture is shown in figure 5 consists of the following parts:

- Moodle LMS that runs on the server, this part is implemented in PHP and supports databases like MySql, Postgreql, Oracle or Microsoft's SQL server.
- Webservices layer, which is a Moodle plug-in our group has developed. This part also runs on the same server as Moodle. These web services implement both the XML-RPC and SOAP standards. However the mobile client uses only the XML-RPC protocol because -theoretically- will be more efficient in this kind of scenario. The analysis of this issue is a material for another eventual research.
- Mobile Client, through the web services layer, Moodlbile synchronizes the data with the Moodle server.

In working online, the mobile client application uses the webservice layer to access the new information from the Moodle server. This new information is sent to the mobile client and stored temporarily for offline access. When the student updates an activity, the changes are stored locally on the mobile device database, and synchronously sent to the Moodle server database (Guerrero, Forment, Gonzalez, & Penalvo, 2009).

On the other hand, the Moodbile client can work offline as well as online. In working online, the user will be able to access the information stored on the mobile device in the last synchronization. The mobile user will also be able to do some update information from the mobile device. This now updates is stored in the mobile's database, and it is ready to be sent on first synchronization (Forment, Guerrero, & Snchez, 2008).

Considering online status, the student can use the mobile device to access very specific information about recent events in short connections or extend the learning process on the move, while he/she can work with: forums, wiki contents, glossary, entries, internal mail messages and calendar.

However, regarding to mobile capabilities, the developers didn't design a full Moodle client that is able to perform all the tasks performed from the web interface. Instead, they considered that the mobile device could be useful to do short connections to the Moodle system accessing specific information with limited updates (Forment, Guerrero, & Snchez, 2008; Guerrero, Forment, Gonzalez, & Penalvo, 2009).

Conclusion

When implementing a mobile learning application, it doesn't matter to adopt a specific theory of learning design, but useful application must be kept under the general frameworks of the pedagogical concepts. Mobile technology is growing dramatically. Near future will overcome the most of its limitation with some exceptions such as screen size and battery life which the developers have to adapt with. Examples presented in this paper give a good practice, and it is represented as a guidance for both application developer and learning resources designers that show how to overcome device limitations with pedagogical consideration.

References

[1] Boggs, R. (2002). ECAR study: Trends in wireless communications in higher education. *Retrieved January, 29*, 2003.

[2] Boja, C., Batagan, L., Mastorakis, N. E., Croitoru, A., Balas, V. E., Son, E., et al. (2009). *Software characteristics of m-learning applications*.

[3] Cavus, N., & Ibrahim, D. (2009). mâ€ےLearning: An experiment in using SMS to support learning new English language words. *British Journal of Educational Technology, 40*(1), 78-91.

[4] Chiu, C. H., Hung, C. M., & Street, S. L. (2009). Evaluation of Applications of Personal Digital Assistants in Elementary Education. *WSEAS Transactions on Advances in Engineering Education*, 443-453.

[5] Cobcroft, R. S., Towers, S. J., Smith, J. E., & Bruns, A. (2006). Mobile learning in review: Opportunities and challenges for learners, teachers, and institutions.

[6] Crow, R., Santos, I. M., LeBaron, J., McFadden, T. A., & Osborne, F. C. (2010). Switching gears: moving from e-learning to m-learning. *MERLOT Journal of Online Learning and Teaching, 6*(1), 268-278.

[7] Dochev, D., & Hristov, I. (2006). Mobile Learning Applications Ubiquitous Characteristics and Technological Solutions. *Cybernetics and Information Technologies, 6*(3), 63-74.

[8] Ferioli, F., & Van der Zwaan, B. C. C. (2009). Learning in times of change: a dynamic explanation for technological progress. *Environmental science & technology, 43*(11), 4002-4008.

[9] Fetaji, B., & Fetaji, M. (2011). *Analyses and review of M-learning feasibility, trends, advantages and drawbacks in the past decade (2000-2010)*.

[10] Fishman, B., Soloway, E., Krajcik, J., Marx, R., & Blumenfeld, P. (2001). Creating scalable and systemic technology innovations for urban education. *Ann Arbor, 1001*, 48109-41259.

[11] Forment, M. A., Guerrero, J. C., & S‹ĺnchez, I. A. (2008). *Moodlbile: extending moodle to the mobile on/offline scenario*.

[12] Gregson, J., & Jordaan, D. (2009). Exploring the challenges and opportunities of m-learning within an international distance education programme. *Mobile Learning*, 215.

[13] Guerrero, M. J. C., Forment, M. A., Gonzalez, M. C., & Penalvo, F. J. G. (2009). *SOA initiatives for eLearning: a Moodle case*.

[14] Harris, P. (2001). Goin'mobile. *Learning Circuits*.

[15] Herrington, A., & Herrington, J. (2007). Authentic mobile learning in higher education.

[16] Herrington, J., Herrington, A., Mantei, J., Olney, I., & Ferry, B. (2009). Using mobile technologies to develop new ways of teaching and learning.

[17] Holzinger, A., Nischelwitzer, A., & Meisenberger, M. (2005). *Mobile phones as a challenge for m-learning: Examples for mobile interactive learning objects (milos)*.

[18] Kennedy, K. (2003). Writing with web logs. *Technology and Learning Magazine, 23*.

[19] Kukulsha-Hulme, A., & Traxler, J. (2007). Design for Mobile and Wireless Technologies in H. Beetham & R. Sharpe (2006) Rethinking Pedagogy for the Digital Age: Routledge, London.

[20] Kukulska-Hulme, A., & Pettit, J. (2009). Practitioners as innovators: Emergent practice in personal mobile teaching, learning, work, and leisure. *Mobile Learning*, 135.

[21] Low, L., & O'Connell, M. (2006). *Learner-centric design of digital mobile learning*.

[22] Maniar, N., Bennett, E., Hand, S., & Allan, G. (2008). The effect of mobile phone screen size on video based learning. *Journal of Software, 3*(4), 51-61.

[23] Motiwalla, L. F. (2007). Mobile learning: A framework and evaluation. *Computers & Education, 49*(3), 581-596.

[24] Park, Y. (2011). A pedagogical framework for mobile learning: Categorizing educational applications of mobile technologies into four types. *The International Review of Research in Open and Distance Learning, 12*(2), 78-102.

[25] Pieri, M., & Diamantini, D. (2009). from E-learning to mobile learning: New opportunities. *Mobile Learning*, 183.

[26] Sharples, M., Taylor, J., & Vavoula, G. (2005). Towards a theory of mobile learning. *Proceedings of mLearn 2005, 1*(1), 1-9.

[27] Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning, 2*(1), 3-10.

[28] Traxler, J. (2009). The evolution of mobile learning. *The evolution of mobile teaching and learning*, 1-14.

[29] Wood, S. L. (2010). Technology for Teaching and Learning: Moodle as a Tool for Higher Education. *International Journal of Teaching and Learning in Higher Education, 22*(3), 299-307.

[30] Yordanova, K. (2007). *Mobile learning and integration of advanced technologies in education*.

N+1 Decision Trees For Attack Graph

Tawfiq S. Barhoom and Lamiya M. EL_Saedi
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
{tbarhoom, lalsaedi}@iugaza.edu.ps

***Abstract*:**

Attack Graph is very useful technique for administrator to map the system vulnerabilities, the information mapped are attack's goals and paths.

In this paper we introduce a novel way to draw an Attack Graph, by using Decision Tree to preprocessing the vulnerabilities information collecting from government institution using NESSUS tool. Decision Tree is a supervised learning classification technique represent paths and text in the nodes and on the edges for verifying the easy understand vulnerabilities. The tree used is very useful in the way of generating the graphs. The graphs are N+1: N for each attribute and one for full graph. This way simplify the way the administrator to learn the situation by minimize the size of graph and then evaluate the system vulnerabilities.

Index Terms: *Data Mining, Decision Tree, Security, Attack Graph, Graph.*

1. INTRODUCTION

Attack graph is a way used by administrator to discovery and analysis network attack models. And used to specify and determine how vulnerable their systems and what security measures to deploy to defined their systems. Attack graphs can be used as a useful tool in several areas: in Network security including (Intrusion detection, defense, and forensic analysis). So, the administrator used attack graph to generate information and to make decisions. *The first* reasons achieved by ask "what attacks is my system vulnerable to " and "how many different ways can an attacker reach a final state to achieve his goal?" *The second* reason achieved by ask "which set of actions should I prevent to ensure the attacker can't achieve his goal?" or

"Which set of security measures should I deploy to ensure the attacker can't be achieved his goal?" The paths in attack graph represent the scenario of attacker to achieve his goal. These paths called actions. [5].

You can use Attack graph as a system helps an administrator to find an answers for some questions like "if

an attacker start from this point what is the goal for him or are there another path can he use it to verify his goal?" [3].

"Attack graph is a tool to analyze multi-stage, multi-host attack scenarios in a network .It is a complete graph where each attack scenario is depicted by an attack graph which is essentially a series of exploits" [4].

To generate an Attack graph is very difficult process for any user because you can't limited and predicted the paths attacker that he uses to verify his goal. Especially in Network, because it's able to be increase and very complicated. So, the first thing is you must understand the security problem clearly to minimize the security problem. Second put automatic solutions according to the security problem and what configurations need to make the modification is easier for user [2].

In this paper we suggest to use Data Mining methods to generate attack graph. Which is a novel way used in every environment of science. Here we apply it in security science to proof that the data mining able to meet your entire request at, via some knowledge for your work, and some knowledge and experience for how to use Rapid Miner to generate your idea. You can work in two areas for learning in Data Mining, supervised learning and unsupervised learning.

There are many stages in Data Mining such as preprocessing, classification, association rule, clustering, outlier and evaluation. For each stage there are many methods and models depends on variant algorithms to apply what you need via drag the icons name of the method on the work area which represent as box titled with name of method you are chose. Then connect these boxes with line from one box to another.

In this paper we represents related works, government institution network structure, basic information about decision tree, working with data mining, actual work, discussions and conclusions.

2. RELATED WORKS

1. Study of Generating Attack Graph based on Privilege Escalation for Computer Networks (2008): in this paper, the authors worked very

hard to generate a general system that associate between the types of vulnerabilities and the kinds of attack that can be occur on those vulnerabilities. They insert all information in five tables by using RDBMS and establishing a relation between these tables to predict the attack graph based on privilege escalations. The tables' names are: Vulnerability Type, Attack type fact, Attack Type Prerequisite, Attack Type Consequence and VulName to PID. Table Vulnerability type predict consists of predicates possibly used to represent the prerequisites and consequents of attacks. Attack types are stored into table Attack Type Fact. Attack Type Prerequisite and attack Type Consequence have the same structure, but are used to keep the prerequisite and consequences of known attack types. VulName to PID is designed to automatically achieve transformation from discovered vulnerabilities to atomic predicates and lays the information for later processes. Also introduce three kinds of attack graph (Network analysis of vulnerability) Model checking, logic programming and exploit-dependency graph search algorithm. [7]

2. Network Security Evaluation through Attack Graph Generation (2009): this paper helps me to identify the attributes that needs to draw the attack graph. These attributes are, " 1- ***Computer and Network*** H= {h1, h2 … hm}, to represent these devices for example, computers, routers, switches. HOSTID is the unique identifier of host on the network, it can be the IP address or host name. OS is the type and version of operation system. SVCS is the list of network service types with respective network port numbers which describes the services on host and the information on service monitor ports. VULS is the host computer vulnerability list which may include the security bug information of installed software or environment misconfigures information, and is presented by its CVE ID.

2- User Privilege:

Privilege class	***Role description***
ROOT	System administrator, managing all system resources.
USER	Any general system user, which is created by administrator.
ACCESS	Remote visitors which may access network services

Table (1): define the role for each privilege class.

3- Connecting Relationship

The Internet is structured based on TCP/IP protocol family

Protocol Layer	***Link Relation Example***
Application Layer	HTTP FTP
Translation Layer	TCP UDP
Network Layer	ICMP
Data Link Layer	ARP

Table (2): define the kinds of protocol in each layer of network.

The connection relations between hosts: ***HSRC*** represents source host. ***HDST*** represents destination host. ***Protocols*** are a sub-set of connection relations sets between source host and destination host. When there is no relation between source host and destination host, *Protocols* is empty set. When the source host is the same as destination host, the connection relation is local connection, at this time, *Protocols = {localhost}*.

4- Attack Rule:

Preconditions				***Postconditions*** **(results set)**		
Src_privilege	***Dst_privilege***	***Vuls***	***Protocols***	***Rslt_privilege***	***Rslt_protocols***	***Rslt_vuls***

Table (3): define the pre-condition and post-condition for each privilege.

Src_privilege represents the lowest privilege which attacker should have on the host where the attacks are launched. ***Dst_privilege*** represents the highest privilege which attacker should have on the object host. ***Vuls*** represents the vulnerability which the attack rule depends on. ***Protocols*** describe the needed connection relation between attack host and object host. ***Rslt_privilege*** describes the privilege which attacker can get on object host after an attack is successfully completed. ***Rslt_protocols*** is the network protocols set which is added by attacks. If the attacked host can use the network protocols in this set to access a host on the network, the current attacking host can get the ability to access this host. If the attack rule doesn't influence the current network connection relations, *Rslt_protocols* will be an empty set. When *Rslt_protocols= {all}*, this represents that the current attacking host can get the attacked host's total ability to access the object network. ***Rslt_vuls*** is the newly added vulnerability set on attacked host after attack is successfully implemented, and it describes the dependent relation between vulnerabilities." [8]

3- Improving Attack Graph visualization through data reduction

And Attack Grouping (2008): in this paper authors represents they own methodology to decrease the size of the attack graph, especially in large enterprise network, by grouping the attack-paths that have the same prevent configuration to solve the problem. Via create the virtual nodes in the model of network topology to include these paths which depends on increase the understandability of data. This technique minimize the size of attack graph to be easy to understand the paths of attacker that be taken to verify his or her goal. The main approach is:

- Developed an algorithm to trimming the paths that not helpful the user to understand the security problem core.

- developed a method to create virtual nodes to represents grouping of similar exploitations.

They use an attack-graph toolkit (MulVAL) and using GraphViz to construct the image and applying clustering technique. [6]

4- Tools for Generating and analyzing Attack Graphs (2004): in this paper the authors represents how to generate attack graph automatically and to analyze system vulnerabilities. The advantage of using attack graph is to evaluate the security of network. Via enough information about the infrastructure of connection (ports, communication, firewall configuration…). This technique actually needs to be updated every time to know if there are other problems, to show if the last problem was solved or there another attack needs higher defense. [5]

5- From Attack graphs to Automated configuration management An Iterative approach (2009): The authors says that the attacker must be understand the Network system for this company to be allow to verify his goal, and in the other side the security persons must be check and predict undirected access point that can attacker use it to access the system network from it and close it and make a suitable configuration change specify that point. The authors idea is present the paths that could attacker use it and which known as useful way or path and close or prevent the useless path. From this idea the attacker can't be use only the useful way. There approaches are:

- using Trimming algorithm: but the problem in this algorithm is can't specify the vulnerabilities. So, the

authors use the SAT solving Techniques to solve the problem automatically by suggest the best modification address the security problem that appear on the attack model.

- They would like to put suggested modification to solve to solve the problem maybe present from attack. So, they use the following way: gives the user ability feedback to SAT solver, then the putting restriction be easy. Every one can be use it, coast deployment, and what can happen if the attack is successful. All of these things can be optimized by unified the framework. Authors transforming attack graphs to Boolean formulas. [2]

6- An Intelligent Technique for generating Minimal attack Graph (2009): the authors used a special purpose search algorithm in artificial intelligent domain for finding out solution within a large state space. They used SGPlan Planner for finding the attack paths.

Initial state, goal state and the state transition operators are provided as input to the planner. They used planner to generate attack paths because 1) It prunes unnecessary actions from the system and finds the shortest path. 2) It allows addition of actions to the plan where ever and wherever they are required. 3) It uses richer input language PDDL (Planning Domain Definition Language) to express complex state space domains relatively easier than custom built analysis engines. 4) It does not suffer from state space explosion problem.

How to generate minimal attack graph by using Planner?

- Minimal attack graph consist only the attack pathes that terminate to a specific goal node.
- Minimal attack graph does not contain a redundant edge or paths. So, it is help the network administrator to take a suitable solution to prevent the new attack to be occurring with different scenario.
- It depends on remove the backtracking from attack graphs and reduces the generation time from exponential to polynomial.
- Planner generates acyclic paths which give a minimal attack graph.

The technique was depending on the initial configuration of the network and the vulnerability analysis. Using PDDL language to write (domain.pddl and fact.pddl). To generate other attack paths,

modify the fact.pddl file. The authors use a case study to explain the above techniques. [4]

7- Analyzing and comparing the protection Quality of Security Enhanced Operating System (2009 or 2008): in this paper the authors introduce the notion of vulnerability surface under attack scenarios as the measurement of protection quality, and implement a tool called VulSAN for computing such vulnerability surface. VulSAN can be used by LINUX system administrators as a system hardening tool, to compute the host attack graphs for attack scenarios. The approach is generates all possible attack paths that can lead an attacker to control of the system. Analyze the QoP (Quality of Protection). Under multiple attack scenarios, which have two approaches: 1) the objective of the attacker (load kernel module or plant a Trojan horse). 2) Initial resources the attacker has (that connect the machine from network, or have a local account). 3) The VulSAN gives all possible attack paths. VulSAN contains the following components: 1) Fact collector: retrieves information the system state and security policy, and encoding the information as facts in prolog. 2) Host Attacker Graph Generator: takes system facts, a library of system rues and the attack scenario as input, and generates the host graph attack. 3) Attack Path Analyzer: find al the minimal attack paths in a host attack graph.

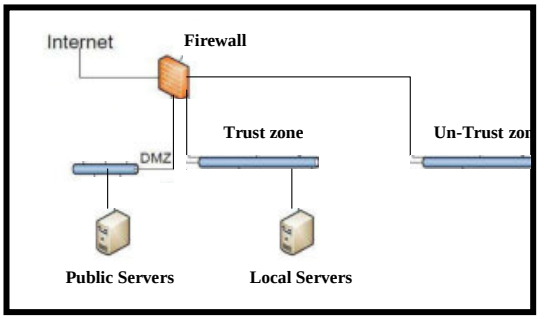
The authors make a comparison between SELinux with AppArmer. They use three scenarios to evaluate there approach: 1) Remote attacker to install a root kit (Assume it is installed by loading a kernel module). 2) Remote attacker to plant a Trojan horse. A) "Strong Trojan case", attacker can create an executable in a folder on the executable search path or user's home directory. B) "Week Trojan case", the attacker can create an executable in any folder such that a normal user process (with a user uid and runs under unconfined domain in SELinux or is not confined by any profile in AppArmor) can execute. For both cases after the Trojan program is executed the process should be unconfined. 3) For a local attacker to install a root kit.

Among the three cases, AppArmor has the smallest vulnerability surface. SELinux has all the minimal attack paths AppArmor has and some addition ones. They found that the SELinux policy in Fedora 8, which is SELinux targeted policy,

offers significantly better protection than the SELinux in Ubuntu 8.04 server edition.

3. Government institysion network structure

The structure of Network Company is, they have one firewall divided into three segments: trust, un-trust and DMZ. All public servers are in DMZ. All local servers in local trust zone. See figure (III.1). Public servers contains mail server, two web servers one of them use LINX operating system and the other used windows server operating system, switch, router, backup server, and DNS. But the local servers contain monitor server, oracle application server, proxy, domain controller, application server, IDS, scanning vulnerability, monitor server, and DB-test. And when aggregate information of vulnerabilities kind with public server's computers to each other, we found that the DNS have variant of medium and low severities possible attacks. And the web server which used LINX has variant high and medium severities possible attacks. But all local servers have low severity possible attacks. Therefore, based on these results we make a decision to work only public servers. Because there are different types of severities.



4. basic information about decision tree

Decision tree is a type of tree-diagram. Which is a common method used to predict the output in data mining.

Consist of internal node is a test on an attribute, branch represents an outcome of the test, and leaf node represent a class label.

There is a large number of decision-tree induction algorithms described primary in the machine learning. To build the tree there are many techniques such as top-down-tree, top-down-induction of decision tree, greedy tree growing, and recursive partitioning.

The strategy: choose attribute that results in greatest information gain. Where (information gain = information before split – information after split).

The knowledge which represents depends on IF-THEN rules. One

rule is created for each path from the root to leaf. The leaf node holds the class prediction. Also, rules are easier to understand.

Decision tree avoid over-fitting by using either pre-pruning or post-pruning. [19].

5. working with data mining

Actually, the previous researches search to best way to draw a graph that allow to represent difficult information in a tree graph, to display it in simple, easy, useful and more understandable way for administrator or for any interested people in this area. So, our choices are working with techniques in Data Mining. Working in data mining is very useful and comfortable. Because you don't need to create a relation between many tables like data base to get result. Or you don't implement any algorithms to draw a graph. Just you need a data set (collect of columns or attributes for any subject you need to analyze it) in one table according for what you are interesting.

We use Rapid Miner 5 to establish an attack graph in very easy way. Without write any coding just use the Decision Tree model which is a classification technique known with supervised teach, for help us to draw a graph. Decision Tree is classification model depends on a target class. Used for classification and decision making. Represent attributes name in nodes in the first and middle levels in a tree. The last level is the content of target class that we need to classify depends on it. The title on the path represents the contents of upper attribute node. It is very easy to read and understand for administrator to trace types of computers, operating system, vulnerability description, severity ratio and solution.

In this paper interesting for vulnerabilities that infect public servers in government institution. We asked this institution to collect some information depends on related work number (2). Data we have obtained are introduced in figure (14). The government institution use NEESUS tool to collect information about vulnerabilities number on each computer connects into institutions network.

Decision tree can draw a graph that illustrates the kinds of servers, operating systems on each computer. Also can draw a graph that represents the types of vulnerabilities that infect each public server and the solution to prevent the attack with high severity. And by using this tool you can see and read everything clearly. And the attack graph can be very simple or complicated. If you need the final full graph, you must insert all attribute in the Decision Tree model.

Now specify and define the attributes used in this work. The attributes are 1- H the kind of computer (computer, router, switch…), 2- HOSTID the host number or the IP address of the

computer, 3- OS the operating system on each computer (windows, LINUX…), 4- SVCS describe the purpose of the computer (web server, mail server, DNS…), 5- VULID the vulnerability number, 6- VULDESC describe the vulnerability number, 7- #OF ISSUES represents how many times NEESUS tool run on the computer (which is not very important but give more details in the graph), 8- SEVERITY describe the dangerous ratio for vulnerability (high, medium and low), and 9- SOLUTION that represent the arbitrary subjected solutions to prevent the attack.

So, we have nine attributes that can give nine sub-attack graphs to easy represent the information for administrators.

Note: N+1 means, N is number of sub-tree or sub-graph which you can draw and comes from number of attributes in data set. 1 is the graph which contains all attributes.

The operation done like this, open Rapid Miner and create new file for new work. Then from file menu import the data file according the type of file you use to store your information. Then choose the repository to include the storage data and connected with the operation defined in the Rapid Miner.

It is important to solve the vulnerability with high and medium severity. If the problem solved must re-again apply the NESSUS tool to verify if the vulnerability is release or not, then update the Excel file where we store the data and re-import the file to Rapid Miner.

6. the actual work

At first import an excel file where data was stored to Rapid Miner. Then select the repository tool and added on the main process. After that we choose the set role operator property to identify the target class by chose the attribute from name and label from target role. Then if you need to simplify the graph you must add a select attribute operator and from property you can chose subset from attribute filter type, then select the attributes from attributes. If your request is the full graph don't add the last operator. Finally, add the decision tree operator to generate a decision tree graph. See figures (1) and (2). After running the process you can see the result graph. In this operator just the related attributes only appear in a decision tree graph. For example see figures (3), (4), (5), (6), (7), (8), (9), (10) and (13).

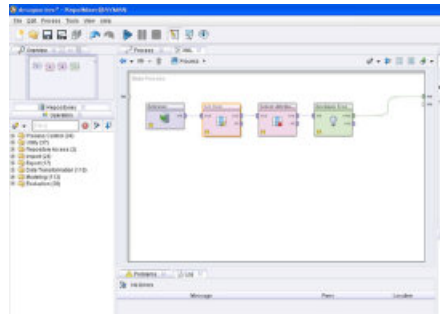


Figure (1): This figure represent the Rapid Miner interface and the process interface.

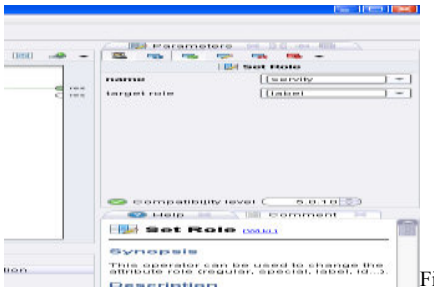


Figure (2): This figure represent the property of set role operator. Where the target class is severity

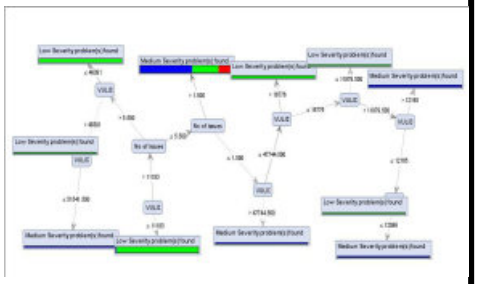


Figure (3): The graph represents the severity ratio as a target class (high, medium, low). The related attribute is VULID and Number of issues.

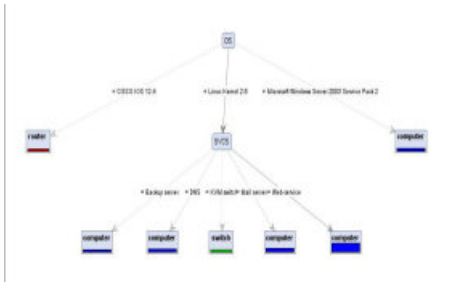


Figure (4): Represent the types of computers in the institution and the operating system on each computer. Where target class is H and related attribute is OS and SVCS.

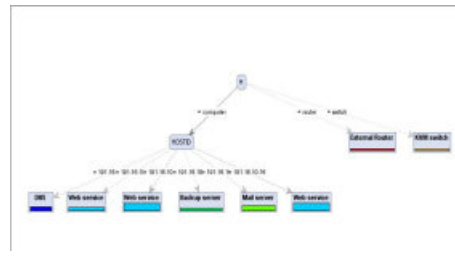


Figure (5): Represent the host description as target class for each HOSTID and related attributes are H and HOSTID.

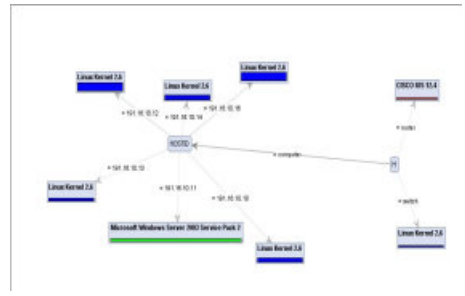


Figure (6): Represent the operating system as target class for each HOSTID and related attributes are H and HOSTID.

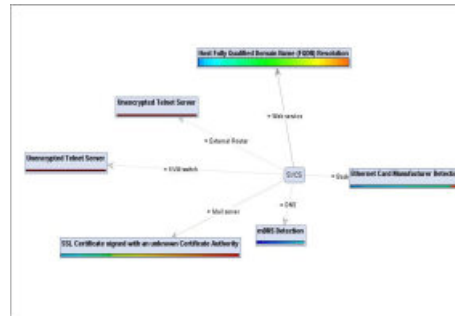


Figure (7): Represent the vulnerability description as target class for each computer and related attribute is SVCS.

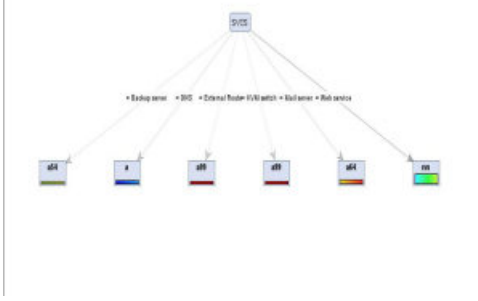


Figure (8): Represent the solution as target class for each computer and related attribute is SVCS.

Figure (9): Represent the vulnerability ID as target class for each computer and related attribute is SVCS.

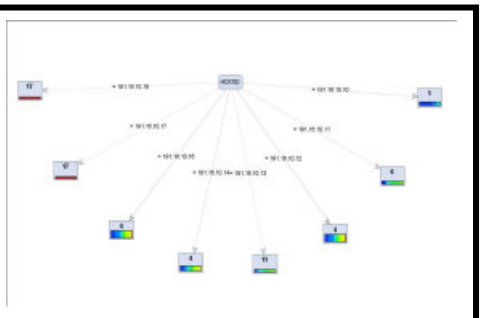


Figure (10): Represent the number of issues as target class on each computer and related attribute is HOSTID.

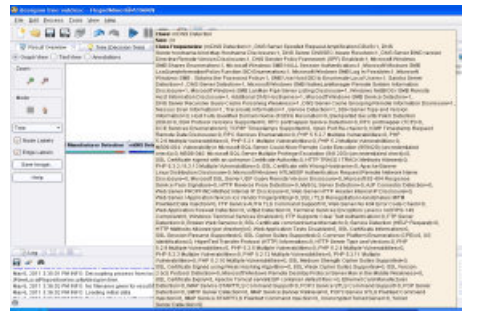


Figure (11): When you point over any node in the last level like m-DNS-detection all information about vulnerability description that infects a specific computer. If the vulnerability equal to one, this means is found else not found.

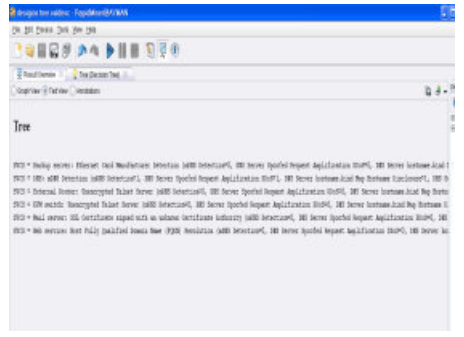


Figure (12): the text view window for vulnerability description graph.

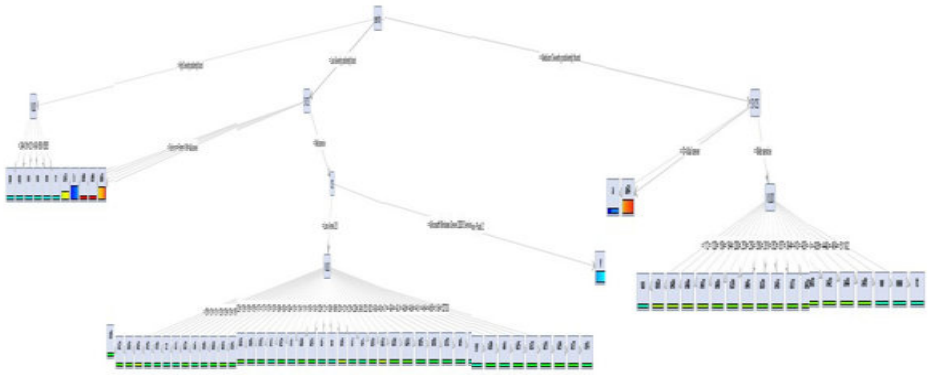


Figure (13): The full attack graph which represent all elements belongs to data set that stored in Excel file. Where the target class is SOLUTION and in the select attribute operation select all attributes without number of issues.

H	HOSTID	OS	SVICS	VULSID	#of issues	Description	Severity
computer	191.16.80.10	Linux Kernel 2.6	DNS	12218	5	mDNS Detection	Medium Severity problem(s) found
computer	191.16.80.10	Linux Kernel 2.6	DNS	35450	1	DNS Server Spoofed Request Amplification DDoS	Low Severity problem(s) found
computer	191.16.80.10	Linux Kernel 2.6	DNS	35371	1	DNS Server hostname.bind Map Hostname Disclosure	Low Severity problem(s) found
computer	191.16.80.10	Linux Kernel 2.6	DNS	35373	1	DNS Server DNSSEC Aware Resolver	Low Severity problem(s) found
computer	191.16.80.11	Microsoft Windows Server 2003 Service Pack 2	Web service	10267	6	SSH Server Type and Version Information	Low Severity problem(s) found
computer	191.16.80.11	Microsoft Windows Server 2003 Service Pack 2	Web service	12053	6	Host Fully Qualified Domain Name (FQDN) Resolution	Low Severity problem(s) found
computer	191.16.80.11	Microsoft Windows Server 2003 Service Pack 2	Web service	38520	6	Backported Security Patch Detection (SSH)	Low Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	24907	4	PHP < 5.2.1 Multiple Vulnerabilities	High Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	31649	4	PHP < 5.3 Multiple Vulnerabilities	High Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	35635	2	MS09-004: Vulnerability in Microsoft SQL Server Could Allow Remote Code Execution (959420) (uncredentialed check)	High Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	15901	1	SSL Certificate Expiry	Medium Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	12085	1	Apache Tomcat Default JSP contained default files	Medium Severity problem(s) found
computer	191.16.80.12	Linux Kernel 2.6	Web service	19506	20	Nessus Scan Information	Low Severity problem(s) found
computer	191.16.80.16	Linux Kernel 2.6	Web service	12085	1	Apache Tomcat Default JSP contained default files	Medium Severity problem(s) found
computer	191.16.80.16	Linux Kernel 2.6	Web service	19506	20	Nessus Scan Information	Low Severity problem(s) found
computer	191.16.80.16	Linux Kernel 2.6	Web service	10287	19	Traceroute Information	Low Severity problem(s) found
computer	191.16.80.16	Linux Kernel 2.6	Web service	22964	70	Service Detection	Low Severity problem(s) found
snvicls	191.16.80.17	Linux Kernel 2.6	KVM snvicls	42263	17	Unencrypted Telnet Server	Low Severity problem(s) found
snvicls	191.16.80.17	Linux Kernel 2.6	KVM snvicls	10281	17	Telnet Server Detection	Low Severity problem(s) found
router	191.16.80.18	CISCO IOS 12.4	External Router	42263	17	Unencrypted Telnet Server	Low Severity problem(s) found
computer	191.16.80.14	Linux Kernel 2.6	Mail server	15901	1	SSL Certificate Expiry	Medium Severity problem(s) found
computer	191.16.80.14	Linux Kernel 2.6	Mail server	52810	1	POP3 Service STLS Plaintext Command Injection	Medium Severity problem(s) found
computer	191.16.80.14	Linux Kernel 2.6	Mail server	52809	1	IMAP Service STARTTLS Plaintext Command Injection	Medium Severity problem(s) found
computer	191.16.80.14	Linux Kernel 2.6	Mail server	19506	20	Nessus Scan Information	Low Severity problem(s) found
computer	191.16.80.14	Linux Kernel 2.6	Mail server	10287	19	Traceroute Information	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	10150	2	Windows NetBIOS / SMB Remote Host Information Disclosure	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	12053	6	Host Fully Qualified Domain Name (FQDN) Resolution	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	10736	12	DCE Services Enumeration	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	25220	11	TCP/IP Timestamps Supported	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	10919	11	Open Port Re-check	Low Severity problem(s) found
computer	191.16.80.13	Linux Kernel 2.6	Backup server	10114	10	ICMP Timestamp Request Remote Date Disclosure	Low Severity problem(s) found

Figure (14): represents the attributes and sample of data set.

All pictures in the above display how you can divide a huge graph to subset graphs each of them concerns with specific information depends on the administrator request and for what information he need to show. But figure (13) display the complete graph for all information stored in excel file.

You can read the decision tree graph from above to down. For example look at figure (7). It is illustrate if the computer is classified as DNS then there are many vulnerabilities found in this computer. You can know there are variant vulnerabilities from the gradient color appearing with box in the leaf. And if you want to know what are these vulnerabilities you can put the mouse on any box in the last level without clicking to see a note list tell you what are the kinds of vulnerability found in a specific computer. See figure (11). Also you can read the texts which is appear in the text view tap in the result window for more understanding see figure (12). If you want to know the solution on each part you can see figure (8). Also you can go to solution box with just pointing to see all possible solution you can chose to apply on your system.

7. Discussion

Honestly, Rapid Miner is very useful but if you need to update your data which we written her by Excel file to see the newest effect on your server. You must updated manually and re-again import the file to Rapid Miner.

You must use real data to feel pleasure and reality of work. Her every data is real (true), only the final attribute (solution). We assumed to try if the solution appears in the graph or not. But you can replace the contents of that field with real information.

This tool helps you to evaluate your work if you need. And helps you to know what the nearest solution is for a new attack if you don’t know what the solution is according tracing the decision tree model.

At first we think to generate attack graph by using association rule with FP-Growth which is unsupervised learning. But the graph was appear in very complicated picture and you can't separate the graph to be clear, and we face a problem to illustrate the text of vulnerability description in association rule. Also you can't read the information in a clear manner. So, we use decision tree model to generate the graph.

The operation for collecting data is very difficult and required experts in the area to gather information in a proper manner.

For working with Data Mining you require some experience in this area that allow you to understand and select the suitable method to apply your work in a proper way.

About evaluation for classification model, you can evaluate the model and show the accuracy via validation operator and by specify the size of training and testing. Actually divided into two parts. The training part equal to 70% and the testing part equal to 30%. But this technique required repeated large information in target class to train a model for expecting the label to a new data row comes without that label. You must try all classification models to choose the two best models with higher accuracy. Then insert the two models to T-Test operator which tell you the best one for expecting in a future.

To concatenate local servers' information to table stored in Excel file. You must add a new column to table in figure (14) name it for example SERVER-KIND, which containing one of two kinds public or local in each row. So, you can add or use all company information in a proper and easy way. And if you need to present a chart especially for new column you insert, you must choose the SERVER-KIND as target class.

8. Conclusions

There is an easy way to generate an attack graph, and released all problems faced by the previous authors.

We suggest Rapid Miner tool which is very easy to obtain and install in your computer, not like the tools used by the above authors.

You can obtain a minimal attack graph by using decision tree model.

You can read and trace the graph in very easy way to understand what happens on your network.

You can see your all elements in one decision tree, or you can divided to sub-trees to minimize the graph depends on your interesting.

Data Mining could be use as a business process.

acknowledgments

We need to thanks and appreciation the government institution in Gaza-Palestine, under the chairmanship of Suhail Madoukh for all helps.

REFERENCES

[1] H.Chen, N.Li, and Z. Mao, Analysing and comparing the Protection Quality of Security Enahanced Operating Syestems. 2009

[2] J. Homer, X. Ou, and M.A. McQueen, From Attack graphs to Automated configuration Management An Iterative approach. 2009.

[3] T. Heberlein, M.Bishop, E. Ceesay, M. Danforth, and C.G.Senthilkumar, A Taxonomy for comparing Attack-Graph Approaches, 2004.

[4] N.Ghosh, and S.K.Ghosh, An Intelligent Technique for generating Minimal attack Graph. 2009.

[5] O.Sheyner,and J.Wing. Tools for Generating and Analysing Attack Graphs, 2004.

[6] J.Homer, A.Varikuti, X.Ou, and M.A.McQueen. Improving Attack Graph Visualization through Data Reduction And Attack Grouping. 2008

[7] X.Chen, J.Li, and S.Zhang. Study of Generating Attack Graph based on Privilege Escalation for Computer Networks. 2008

[8] Z. Lufeng, T.Hong, C. YiMing, Z. JianBo. Network Security Evaluation through Attack Graph Generation. 2009

[9] Rattikorn H. t and Phongphun K. , Host-Centric Model Checking for Network Vulnerability Analysis, 2008, Annual Computer Security Applications Conference

[10] Diptikalyan S., Extending Logical Attack Graphs for Efficient Vulnerability Analysis, 2008.

[11] Yinqian Z., Xun F., Yijun W., Zhi X., Attack Grammar: A New Approach to Modeling and Analyzing Network Attack Sequences, 2008, Annual Computer Security Applications Conference.

[12] Scott O'H., Steven N., and Kenneth P., A Graph-Theoretic Visualization Approach to Network Risk Analysis, 2008.

[13] Somak Bhattacharya, S. K. Ghosh, An Attack Graph Based Risk Management Approach of an Enterprise LAN, 2008, Journal of Information Assurance and Security 2, pp 119-127.

[14] S.M. Welberg, Vulnerability management tools for COTS software - A comparison, 2008.

[15] ASHOK R. V., VISUALIZATION TECHNIQUES IN ATTACK GRAPHS, 2009, Report for Master Research.

[16] Kyle I., Richard L., Keith P., Practical Attack Graph Generation for Network Defense, 2006.

[17] Lingyu W., Anyi L., Sushil J., Using attack graphs for correlating, hypothesizing, and predicting intrusion alerts, 2006, Computer Communications, pp 2917–2933.

[18] Steven N., Sushil J., Managing Attack Graph Complexity Through Visual Hierarchical Aggregation, 2004.

[19] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2001. The Morgan Kaufmann.

Conference Program

تحت رعاية

أ.د. يونس عمرو - رئيس الجامعة

تنظم كلية التكنولوجيا والعلوم التطبيقية في جامعة القدس المفتوحة

مؤتمراً بعنوان

"المؤتمر الدولي لتكنولوجيا المعلومات والاتصالات - التقنيات و التطبيقات"

وذلك صباح يوم الثلاثاء الموافق 2012/6/26 الساعة التاسعة والنصف صباحاً في قاعة ليدرز LEADERS الماصيون ،رام الله

وعبر الفيديو كونفرنس مع قطاع غزة في قاعة المؤتمرات/فرع غزة/النصر

التسجيل: 9:30 –10:00	
عريف جلسة الافتتاح : أ .لوسي حشمة / مديرة دائرة العلاقات العامة	
10:00 – 10:45 جلسة الافتتاح (القاعة 1) النشيد الوطني الفلسطيني كلمة رئيس الجامعة أ .د .يونس عمرو كلمة د. صبري صيدم مستشار الرئيس لشؤون التكنولوجيا كلمة وزير التعليم العالي كلمة وزير تكنولوجيا المعلومات والاتصالات كلمة مندوب شركة جوال كلمة اللجنة التحضيرية للمؤتمر	
استراحة: 10:45 – 11:15	
الجلسة الأولى: (11:15 - 12:15)	
القاعة (1) ICT in Business & Education رئيس الجلسة: د. رشيد الجيوسي / جامعة القدس	القاعة (2) Artificial Intelligence, Data Mining & Software Engineering رئيس الجلسة : د. معاذ صبحة/الجامعة العربية الأمريكية
Effects of using Web 2.0 applications in the E-Business course on Palestinian - Nadira Alaraj	Electric Power Load Short Term Forecasting- Ra'ed Basbous
Mobile Learning Applications - Ramy I. R. Ashour (Gaza)	Academic Researcher Information Extraction from the WEB (ARIEW) - Yousef Abuzir and Sodous Kitane
Critical Factors Influencing the Acceptance and Diffusion of e-Government - Mohammed Ayoub	Improving Software Quality Through Requirements Elicitation - Sereen Abu Aisheh
Protection the Copyright in e-Education Process - Osama Marie	A Comparative Study of Statistical and Data Mining Algorithms for Prediction Performance - Amjad Harb
استراحة 12:15 –12:40	

الجلسة الثانية: (12:40 - 2:00)	
القاعة (1) Mobile Networking & Simulation رئيس الجلسة: مراد ابو صبيح / جامعة البوليتكنك	القاعة (2) Databases & Computer Architecture رئيس الجلسة : د. يوسف حسونة/ جامعة بيرزيت
N+1 Decision Trees For Attack Graph - Tawfiq S. Barhoom (Gaza)	Developing New Methods to Find The Number Of RAM Chips In The Memory Decoding To Construct The Required Memory Size - Mohammad Abu Omar
Runtime Replica Consistency Mechanism For Cloud Data Storage - Mohammed Radi (Gaza)	Selectivity Estimation Technique for Wikipedia - Muath Alrammal
Possibility of Applying Green Communications in Palestinian Cellular Networks - Murad Abusubaih	Applying Data Mining Technology in Modeling and Predicting Number of Students in Bedia Center - Ola Rayyan
Comparison Study of Adhoc Networks Routing Protocols Using NS2 - Ola Sbihat	Cell Phone Jamming Device - Rana Alia
استراحة غداء: 2:00 - 3:10	
جلسة الختام : 3:10 - 3:30 (القاعة 1)	

Conference Committees

Organizing Committee

Yousef Abuzir (Organizing Committee chair)
Hasan Silwadi
Islam Amro
Imad Hodali
Sami Zawahreh
Imad Nazzal
Nael Abu-halaweh
Osama Maraie
Shadia Makhlouf
Lucy Heshmi
Waleed Thwieb
Mohammed Iklil
Ibrahim Deliq
Khader Rajabi
Abdelmoem Maraqa
Musa Abu sharar
Fadi Naser
Wisam Saleibi

Scientific Committee

Yousef Abuzir	(Palestine) (Scientific Committee chair)
Emad Nazzal	(Palestine)
Rashid Jayousi	(Palestine)
Osama Marie	(Palestine)
Ibrahim Al Deleq	(Palestine)
Samer Mayaleh	(Palestine)
Bahjat Qazaz	(Palestine)
Murad Abu Sbaieh	(Palestine)
Raid Zaghal	(Palestine)
Nael Salman	(Palestine)
Nael Abu-halaweh	(Palestine)
Khaled Samara	(USA)
Mutamed Khatib	(Palestine)
Khamis Omar	(Jordan)
Bassam Hasan	(USA)
Waleed Salama	(Jordan)
Saleh Abu Assoud	(Jordan)
Mehmet R. Tolun	(Turkey)
Khitam Azaiza	(USA)
Muath ALRAMMAL	(France)
Muath Sabha	(Palestine)
Shawkat Ali A B M	(Australia)
Mohamed Dweib	(Palestine)
Ismail Romi	(Palestine)


ICICT
جامعة القدس المفتوحة
تحت رعاية أ.د. يونس عمرو رئيس الجامعة
تنظم جامعة القدس المفتوحة
كلية التكنولوجيا والعلوم التطبيقية
مؤتمر

قائمة بأسماء أعضاء اللجنة العلمية

الرقم	اللجنة العلمية	البلد
١.	د.يوسف ابوزر (رئيس اللجنة)	فلسطين
٢.	د. خالد سمارة	الولايات المتحدة
٣.	د.معتم خطيب	فلسطين
٤.	د. خميس عمر	الاردن
٥.	بسام حسان	الولايات المتحدة
٦.	د. صالح ابوسعود	الاردن
٧.	د. محمد تولون	تركيا
٨.	د. ختام عزيزة	الولايات المتحدة
٩.	د. بهجت قزاز	فلسطين
١٠.	د. رشيد جيوسي	فلسطين
١١.	د. اسماعيل رومي	فلسطين
١٢.	د. مراد ابوصبيح	فلسطين
١٣.	د. معاذ الرمال	فرنسا
١٤.	د. معاذ صبحة	فلسطين
١٥.	د. سامر ميالة	فلسطين
١٦.	د. احمد شوكت	استراليا
١٧.	د.نائل ابو حلاوة	فلسطين
١٨.	د.رائد الزغل	فلسطين
١٩.	د.نائل سلمان	فلسطين
٢٠.	د.توفيق برهوم	فلسطين
٢١.	د.محمد ذويب	فلسطين
٢٢.	د.اسامة مرعي	فلسطين
٢٣.	م. ابراهيم الدلق	فلسطين
٢٤.	د.وليد سلعوس	فلسطين
٢٥.	د.ايمن وزوز	فلسطين
٢٦.	د. عماد نزال	فلسطين

قائمة بأسماء أعضاء اللجنة التحضيرية

١.	أ.د. حسن السلوادي	عميد برنامج البحث العلمي والدراسات العليا
٢.	م. عماد الهودلي	مساعد الرئيس لشؤون التكنولوجيا والإنتاج.
٣.	د. يوسف أبو زر (رئيس اللجنة)	عميد كلية التكنولوجيا والعلوم التطبيقية.
٤.	أ. سامي زواهرة	عضو هيئة تدريس / فرع بيت لحم
٥.	د. عماد نزال	مدير فرع جنين
٦.	د. أسامة مرعي	عضو هيئة تدريس / فرع نابلس
٧.	أ. شادية مخلوف	مديرة دائرة الجودة.
٨.	أ. لوسي حشمة	مديرة دائرة العلاقات العامة.
٩.	م. وليد ذويب	مساعد عميد كلية التكنولوجيا والعلوم التطبيقية.
١٠.	م. محمد إخليل	عضو هيئة تدريس / فرع بيت لحم
١١.	م. إبراهيم الدلق	عضو هيئة تدريس / فرع طولكرم
١٢.	أ. خضر الرجبي	عضو هيئة تدريس / فرع القدس
١٣.	أ. عبد المنعم مرقة	عضو هيئة تدريس / فرع الخليل
١٤.	أ. موسى أبو شرار	عضو هيئة تدريس / فرع دورا
١٥.	أ. فادي ناصر	مساعد مدير مركز ictc قطاع غزة
١٦.	أ. وسام الصليبي	مسؤول العلاقات العامة في قطاع غزة
١٧.	د. نائل ابو حلاوة	رئيس قسم انظمة المعلومات الحاسوبية
١٨.	د اسلام عمرو	رئيس قسم تكنولوجيا المعلومات والاتصالات

جامعة القدس المفتوحة
كلية التكنولوجيا والعلوم التطبيقية
technology@qou.edu

المؤتمر الدولي
لتكنولوجيا المعلومات والاتصالات
(ICICT'2012)

والذي تنظمه
كلية التكنولوجيا و العلوم التطبيقية
في جامعة القدس المفتوحة

يعدّ المؤتمر الدولي لتكنولوجيا المعلومات والاتصالات (ICICT 2012)الذي تنظمه جامعة القدس المفتوحة في فلسطين المؤتمر الأول من حيث التركيز على عرض أحدث التقنيات المتعلقة بالمعلومات والاتصالات الرقمية .

ويعكس هذا التركيز اهتماماً متزايداً ومتنامياً بكل ما يُستجد من موضوعات أو مجالات تتعلق بتكنولوجيا المعلومات والاتصالات، التي تشكل الموضوع الرئيس للمؤتمر مع ما ينبثق عنه من محاور فرعية نضعها بين أيدي الدارسين والباحثين لدراستها ومعالجتها .

من هنا تم التوجه إلى الأخوة والزملاء الباحثين في حقل تكنولوجيا المعلومات والاتصالات لتقديم إسهاماتهم البحثية حول أي من هذه الموضوعات لعرضها على اللجنة العلمية للمؤتمر تمهيداً لعرضها ضمن فعاليات المؤتمر، حيث ستنشر طائفة مختارة من هذه الأبحاث في عدد خاص من مجلة جامعة القدس المفتوحة للأبحاث والدراسات المحكمة بعد تحكيمها.

أهداف المؤتمر

أهم الأهداف الرئيسة للمؤتمر الدولي لتكنولوجيا المعلومات والاتصالات(ICICT 2012) فهي كما يلي:

- إبراز واقع تعليم التكنولوجيا والمعلومات والاتصالات في مراحل التعليم المختلفة على الصعيد المحلي والإقليمي والدولي ومدى ملائمته لحاجات السوق .
- إبراز دور تكنولوجيا المعلومات والاتصالات في النهوض المجتمع ومؤسساته ومواكبة التطورات الحديثة والسريعة في هذا المجال .
- تقوية وتعزيز العلاقة بين المؤسسات الأكاديمية مع بعضها البعض وبين القطاع العام والخاص .
- معايير الجودة ببرامج وتخصصات تكنولوجيا المعلومات والاتصالات وصناعة البرمجيات.
- تشجيع البحث العلمي في مجال تكنولوجيا المعلومات والاتصالات

الموضوعات التي تم تقديم الأبحاث فيها

- الذكاء الاصطناعي والنظم الخبيرة.
- أنظمة المعلومات الطبية.
- نمذجة الحوسبة.
- نظم الاسترجاع في الكوارث.
- التنقيب في البيانات ومستودعات البيانات.
- الأنظمة الموزعة والمتوازية.
- تكنولوجيا المعلومات والاتصالات في الحكومة
- تكنولوجيا المعلومات والاتصالات في التعليم.
- تكنولوجيا المعلومات والاتصالات في الصناعة.
- الأعمال الإلكترونية.
- حوسبة الإنترنت.
- تكنولوجيا المعلومات والمجتمع.
- قواعد البيانات.
- شبكات الهاتف المحمول وتطبيقاتها.
- نظم الوسائط المتعددة وتطبيقاتها.
- الشبكات ونظم التشغيل.
- التعرف على الأنماط.
- معالجة الإشارات الرقمية.
- محاكاة التكنولوجيا.
- هندسة البرمجيات.
- قواعد المعرفة.
- معالجة اللغة العربية.
- هيكلية الحاسوب.
- هندسة الويب.
- أمن المعلومات

وقد تم عرض عدداً من الأوراق البحثية تناولت مختلف محاور المؤتمر، وقد أوصت اللجنة التحضيرية بالتالي:-

- **أولاً**: العمل على تنظيم مؤتمر في العام القادم بالتعاون مع الجامعات الفلسطينية والعربية.
- **ثانياً**: تعميق التعاون البحثي بين الجامعات الفلسطينية والعربية من خلال عمل أبحاث مشتركة بين أعضاء الهيئة التدريسية في الجامعات الفلسطينية والعربية.
- **ثالثاً**: تشكيل لجنة عليا تضم ممثلين من الجامعات الرسمية والشركات لتحديد أولويات البحث العلمي في مجال تكنولوجيا المعلومات والاتصالات.

وقبل الختام .... تتقدم جامعة القدس المفتوحة ممثلة برئيسها وطاقمها الأكاديمي والإداري بالشكر الجزيل لجميع المشاركين في المؤتمر وتخص بالذكر الباحثين الذين قدموا أوراقا بحثية .كما تتقدم رئاسة المؤتمر ممثلة بالأستاذ الدكتور يونس عمرو بجزيل الشكر للجنة التحضيرية ورئيسها واللجنة العلمية وجميع العاملين في إدارة المؤتمر لجهودهم التي بذلوها من اجل إنجاحه، كما تتقدم رئاسة المؤتمر بالشكر الجزيل إلى شركة الاتصالات الخلوية الفلسطينية (جوال) لتمويلها لهذا المؤتمر، والشكر الخاص لكل وسائل الإعلام التي غطت هذا الحدث الفلسطيني المميز بكل أبعاده وساهمت بإبرازه بالشكل اللائق.

والسلام عليكم ورحمة الله وبركاته

رئيس اللجنة التحضيرية للمؤتمر

د. يوسف أبو زر

بسم الله الرحمن الرحيم

البيان الختامي والتوصيات

للمؤتمر الدولي لتكنولوجيا المعلومات والاتصالات
(التقنيات والتطبيقات)

إدراكاً من جامعة القدس المفتوحة للتطورات التقنية في أنظمة المعلومات والاتصالات، وانطلاقاً من رسالة جامعة القدس المفتوحة وترجمة لرؤيتها وأهدافها في تهيئة الإنسان الفلسطيني وتكوينه أكاديميا ووطنياً وتنمية جميع جوانب شخصيته وقدراته العلمية لتمكينه من مواجهة التحديات ومواكبة التطورات التكنولوجية المتسارعة نظمت كلية التكنولوجيا والعلوم التطبيقية في الجامعة مؤتمرها الدولي الأول لتكنولوجيا المعلومات والاتصالات (التقنيات والتطبيقات) في مدينة رام الله تحت رعاية الأستاذ الدكتور يونس عمرو رئيس الجامعة وبدعم وتمويل من شركة الاتصالات الخلوية الفلسطينية (جوال)، وقد تلقت اللجنة التحضيرية أكثر من خمسة وخمسين ورقة بحثية ومقالة علمية من باحثين من فلسطين والعالم العربي وتم تحكيم الأوراق البحثية والمقالات العلمية المقدمة للمؤتمر من قبل ذوي الاختصاص حيث تم قبول (١٨) بحثاً من أصل (٥٥) للعرض في هذا المؤتمر، **وتركزت البحوث المقدمة في أربعة مجالات ومحاور بحثية هي:-**

- تكنولوجيا المعلومات والاتصالات في الأعمال والتعليم والصناعة.
- الذكاء الاصطناعي والنظم الخبيرة والتنقيب في البيانات ومستودعات البيانات.
- شبكات الهاتف المحمول ومحاكاة التكنولوجيا.
- هيكلية الحاسوب وأنظمة قواعد البيانات .

دوائِرِها و فروعها التعليمية، واخص بالذكر رئيس الجامعة، ونوابهِ الأفاضل، ودائرة العلاقات العامة، ومركز تكنولوجيا المعلومات والاتصالات وعمادة البحث العلمي على ما قدموهُ من جهد لإنجاح هذا المؤتمر. والشكر موصول للداعم الرئيس للمؤتمر شركة الاتصالات الفلسطينية الخلوية جوال على ما قدموه من دعم لإنجاح هذا المؤتمر. وأتوجه إليكم أنتم أيها الأخوة والأخوات الضيوف الأحبة مرحباً بِكُم واشكر لكم تفضلكم بالحضور.

والسلام عليكم ورحمة اللّه وبركاته

الاخوات والأخوة الضيوف،

أيها الحضور الكريم،

يعدُّ هذا المؤتمر الأول من نوعهِ الذي تنظمه جامعة القدس المفتوحة، والذي يركز على عرض أحدث التقنيات المتعلقة بالمعلومات والاتصالات الرقمية. كما وَيبدي اهتماماً متزايداً ومتنامياً بكلِ ما يَستَجِدُ من موضوعات أو مجالات تتعلق بتكنولوجيا المعلومات والاتصالات، التي تشكل الموضوع الرئيس للمؤتمر مع ما ينبثقُ عنه من محاور فرعية نَضعُها بين أيدي الباحثين والمهتمين لدراستها ومناقشتها.

لقد تقدم لهذا المؤتمر خمسة وخمسون بحثا علميا ، قَبِلت اللجنة العلمية منها ثمانيةَ عشر بحثا بعد تَحكيمها من قِبَلِ مختصين في هذا المجال، واعتذر بعضُ الباحثين من العالم العربي عن المشاركة ، وقد تم توزيعُ جلسات المؤتمر على فترتين ، شَمِلت الأولى جلستين في قاعتين بشكل متزامن ؛ حيث سَتَنعَقدُ الجلسةُ الأولى بعنوان (ICT in Bus-ness and education) في القاعة رقم (١) و سَيتمُ من خلالها عرضُ ٤ أربعةَ بحوث علمية برئاسة الدكتور رشيد الجيوسي من جامعة القدس، أما الجلسة الثانية بعنوان Artificial Intelligence, Data Mining and Software engineering فسَتُعقَدُ في القاعة رقم (٢) وسَيتمُ فيها عرض ٤ أربعةَ بحوث علمية برئاسة الدكتور معاذ صبحة من الجامعة العربية الأمريكية.

أما في الفترة الثانية ؛ فسَيتمُ عقد الجلسة الأولى : بعنوان Mobile networking and simulation في قاعة رقم (١) حيث سَيتمُ عرض ٤ أربعةَ بحوث علمية برئاسة الدكتور مراد ابو صبيح جامعة بوليتكنك فلسطين،و في القاعة رقم(٢) ستعقد الجلسة الثانية وهي بعنوان Data base and computer architecture وسَيتمُ عرض ٤ أربعةَ بحوث علمية برئاسة الدكتور يوسف حسونة من جامعة بيرزيت.

وختاماً أتقدمُ بجزيل الشكر إلى زملائي أعضاء اللجنتين التحضيرية و العلمية من داخل الوطن ومن خارجه، والى الباحثين وإلى طواقم جامعة القدس المفتوحة بمختلف

بسم الله الرحمن الرحيم

كلمة رئيس اللجنة التحضيرية للمؤتمر
د. يوسف أبو زر

الأخ أ.د. يونس عمرو / رئيس الجامعة

الأخ د. فاهوم الشلبي / وكيل وزارة التعليم العالي

الأخ د. صبري صيدم / مستشار الرئيس لشؤون التكنولوجيا

الاخوات والأخوة الباحثون

الاخوات والأخوة الضيوف

الحضور الكريم

مع حفظ الالقاب والمسميات

أسعدَ الله صَباحَكم في هذا اليوم من أيامِ جامعة القدس المفتوحة، وأرحبُ بكم أجمل ترحيب في مؤتمر تكنولوجيا المعلوماتِ والاتصالات "التقنيات والتطبيقات" الذي ينعقدُ لأول مرة في فلسطين تحت رعاية أ.د. يونس عمرو/ رئيس الجامعة الذي واكب هذا المؤتمر منذ البدايات ومن فِكرتهِ الأولى حتى أصبح حقيقة واقعة، فله كلُ الشكرِ والتقدير والعرفان، وأشكرُ أستاذنا الكبير الأستاذ الدكتور سفيان كمال على توجيهاتِه التي كان لها الأثرُ الأكبرُ في الوصول بالمؤتمر إلى هذه المرحَلة. كما أرحب بِكُم في هذا المؤتمر الذي تُنظمة كلية التكنولوجيا والعلوم التطبيقية في جامعة القدس المفتوحة، جامعة منظمة التحرير جامعة الحداثة والمستقبل جامعة الشعب الفلسطيني . وأشكر لكم حضورَكمُ و مشارَكتِكمُ لنا هذا المؤتمر.

لقد كان قراراً إستراتيجياً لنا أن يحظى التعليم بالأولوية في توجهنا التنموي المجتمعي المتجدد؛ لأننا نبني على مشاريع تنموية نُفذت على مدار السنوات الماضية من خلال مسؤوليتنا الاجتماعية التي حظيت فيها مشاريع قطاع التعليم بأهم الأولويات. الشباب بشكل عام والطلبة بشكل خاص هم محوُر اهتمامنا، وجلُّ عنايتنا فشركة جوال تقوم على سواعد الشباب الفلسطيني، كما أن الشباب يشكلون النصف الأكبر من المجتمع، كما أننا نؤمن بأن الشباب هم الطاقة المتحركة لتقدم المجتمع والفئة القادرة على تطويره نحو الأفضل، فإن تطوير الشباب وصقل قدراتهم هو تنمية للمجتمع بأسره وضمان لمستقبل واعد.

وكما يعلم الجميع، فإن مجال عمل جوال هو التكنولوجيا والاتصال، وأن الهندسة بفروعها المختلفة تقوم على تطوير هذا القطاع المهم في مجال عمل الشركة.

لذلك كانت جوال دائماً بجانب الطلبة والعمل على رفع مستوى كفاءتهم، وذلك من خلال إخضاعهم للتدريب، ومن البرامج التدريبية التي تقوم جوال بتنفيذها بالتعاون مع الجامعات الفلسطينية والهادفة لتدريب طلبة كلية التجارة الإدارية والتسويقية ضمن برنامج أنا جوال، بالإضافة إلى برنامج تدريبي خاص بطلبة الهندسة من خلال إخضاعهم لتدريب عملي يحاكي الواقع في أقسام الشركة ومحطات البث الخاصة بها، بالإضافة إلى برنامج تدريب للطلبة الخريجين لموظفي جوال.

ختاماً أشكر القائمين على هذا المؤتمر وإدارة جامعة القدس المفتوحة لاهتمامها بالمؤتمرات العلمية، هذه الجامعة التي نشهد لها بتعاونها المميز مع مسيرة مجموعة الاتصالات الفلسطينية وإتاحة المجال لنا لنكون شركاء لهم في جميع المؤتمرات العلمية التي تنظمها. وشكراً لكم.

والسلام عليكم ورحمة الله وبركاته.

بسم الله الرحمن الرحيم

كلمة ممثل شركة جوال، الداعم للمؤتمر
المهندس/ماهر بروق

نجتمع مجدداً اليوم لنجدد التعاون المثمر بين مؤسسات الوطن التي تعكس صورة حضارية، ونحن نؤكد أننا حريصون بعلاقتنا وتعاوننا على الوصول بالوطن إلى مراتب متقدمة. على الرغم من الأحداث والظروف الاستثنائية التي نمرُّ بها وتعترض سبيلنا.

كما نتشرف في شركة جوال بتقديم الرعاية للمؤتمر الدولي لتكنولوجيا المعلومات والاتصالات تحت شعار "التقنيات والتطبيقات"، وذلك إيماناً منا بالدور الذي تنهض به التكنولوجيا في دعم الاقتصاد الفلسطيني وتطويره وبخاصة في ظل التقدم التكنولوجي الدائم.

شركة جوال بدورها تعمل دائماً على مواكبة هذا التقدم التكنولوجي، وبخاصة في مجال الاتصالات، وتأهيل كادر وظيفي قادر على التفاعل مع أحدث التقنيات وتوفيرها للمشتركين.

حيث اعتمدت جوال منذ بداية نشأتها على الكفاءات الشبابية للارتقاء بمستوى خدماتها، إن مشاركتنا ومهمتنا في تنمية المجتمع وخلق مشاريع جديدة ستبقى متواصلة، وهانحن اليوم نبادر من جديد لنكمل العملِ التنموي الذي بدأناه سابقاً، ونبني على نتاج مسيرة طويلة في دعم التعليم منذ سنوات خلت، وسنتمكن من المضي قدماً في طرح سلسلة جديدة من الأفكار الجريئة والطموحة في ظل ابتكار وإبداع لأسس تنموية مستدامة وفريدة، تحقق جزءاً من طموحات أبناء هذا البلد المعطاء.

اليوم أصبحت معلومة عالمية، أي واحد منا يستطيع الوصول إليها، وأصبحت ميزة العصر الذي نعيش فيه الآن أنه عصر اقتصاد المعرفة، والمعرفة هي العتبة الأولى نحو الإبداع والاختراع، فالعصر القادم هو عصر الإبداع والاختراع والذي به نطمح لتحقيق التنمية المستدامة، من هنا أهنئ كل جامعة فلسطينية تمكنت من فتح تخصص منفرد في هذا العلم وأهنئ جامعة القدس المفتوحة التي تمكنت من فتح كلية للتكنولوجيا، ولمؤتمرها هذا أتمنى النجاح والتوفيق.

والسلام عليكم

بسم الله الرحمن الرحيم

كلمة وزارة التربية والتعليم العالي
د. فاهوم الشلبي

إنه لشرف عظيم لي أن أكون معكم اليوم ممثلاُ لمعالي وزير التعليم العالي د. علي الجرباوي، ومشاركاً في هذا المؤتمر العلمي الذي تنظمه وتحتضنه جامعة القدس المفتوحة. حيث يلتقي في هذا الصرح كوكبة من العلماء والباحثين والمهتمين للتدارس والتباحث في موضوع حيوي بالغ الأهمية لوزارة التعليم العالي، ألا وهو تطبيقات تكنولوجيا المعلومات والاتصالات في المجالات المختلفة الحياتية التجارة، الاقتصاد، الهندسة وغيرها.

إن علم تكنولوجيا المعلومات والاتصالات يختلف عن غيره من العلوم والفروع العلمية المختلفة؛ فهو ينمو ويتطور ويتجدد في تسارع عالٍ جداً ليس له سقف، فالذي كنا نتمناه ونعتبره خيالاً علمياً قبل عشر سنين، أصبح اليوم واقعاً يحلم به الصغير قبل الكبير، والذي نحلم به اليوم سوف يصبح واقعاً للأجيال القادمة، وهذا كله بسبب علم التكنولوجيا والمعلومات، هذا التسارع غير موجود في بقية العلوم الأخرى كما هو موجود في علم تكنولوجيا المعلومات والاتصالات، والسبب في ذلك هو حاجة الناس إلى هذا العلم، الحاجة هي التي تولد رغبة، وتولد التفاعل مع هذا العلم، فالطالب وعضو هيئة التدريس، وكذلك الموظف والتاجر والباحث والمؤسسة الخدمية الحكومية أو الأهلية بحاجة إلى الانترنت للحصول على ما يريدون والتعامل مع غيرهم.

سؤالي الآن هو: هل هذا العلم مستقل أو منعزل عن غيره ويطور نفسه بعيداً عن بقية العلوم دون التأثير فيها؟ والجواب لا إنه ليس أنانياً، إنه أساس ويشكّل بذلك أساساً لتطور العلوم الأخرى، وهو أداة طبيعية يفيد منها كل صاحب علم وفن، وأن المعلومة

غداً عندنا مؤتمر في الحوكمة الإلكترونية، وسندعو الأخوة الوزراء حتى يحاكمهم الناس على ما قدموه، أو يبارك لهم الحضور على إنجازاتهم، ولكن التعليم الإلكتروني، وهذا موضع اهتمامنا، ففي لقائنا صباحاً مع د. يونس ود. فاهوم أكدت أنه لا يعقل أن تقوم جامعات عريقة بمستوى كامبريدج واكسفورد وهارفود بتبني التعليم الإلكتروني وتعتمد شهاداته، ولا نستطيع في فلسطين أن نجد آلية لاعتماد هذه الشهادات.

من المعيب –والله– أن نكون تحت احتلال وحصار وتكلفة مالية هائلة، ثم يأتيك طالب يحمل شهادة في التعلم الإلكتروني فترفض شهادته ولا تعتمد بذرائع واهية، والسؤال البسيط الذي ما زال حتى اليوم برسم الإجابة هو: هل من الممكن أن يخضع هذا الطالب لامتحان كفاءة تقيم قدراته وليخضع لامتحان مهما كانت مدته حتى نقول إنه يستحق هذه الشهادة أم لا يستحقها؟ هذا السؤال يجب أن نجيب عليه.

آمل السنة القادمة أن نأتي ومعنا وزير التعليم العالي، ومن خلال هذه المنصة أن نعلن عن ولادة التعليم الإلكتروني كما نراه ويراه العالم، ونحن لدينا الكثير مما نعطيه ونستطيع أن نعطي وأن نقدم.

حياكم الله وبوركتم والسلام عليكم.

بجانبهم وزير أو أمير، بل أبدعوا بإمكاناتهم وبقدراتهم وبكفاءاتهم وبإصرارهم على أن تستمر المسيرة دائماً، والله معنا دائماً بإذن الله، لإننا شعب يستحق الحياة، ولا يستحق القسمة، ولا يقبلها أمام الاحتلال.

أدعو الله أن يكون انطلاق هذا المؤتمر في السنة القادمة في قطاع غزة، بمعية زملائنا وأحبتنا، وأنا أعدكم إن شاء الله أننا سنكون في قطاع غزة خلال الشهر المقبل إن أذن لي د. يونس أن أذهب بصحبته حتى نتابع معاً مسيرة التعليم التي لا ينبغي أن تكون عرضة لأهواء الساسة، أو خاضعة لملفات الانقسام وخلافه.

أما في ما يتعلق بموضوع أنظمة المعلومات، فأنا أردت أن أسرد مجموعة من الأنشطة لأبين أن الأجندة بحمدلله ممتلئة، ويتسارع تنفيذها بحيث لا يزيد الفارق عن نصف ساعة بين نشاط وآخر. والذين لا يعرفون، فإنه خلافاً لكل التقنيات التي حصلنا عليها في الماضي من التلفاز إلى المذياع فإن التكنولوجيا بأجهزتها المختلفة فتحت أمامنا أبواباً على عالم مختلف سريع التحرك، فمن لحق به فاز وارتقى، ومن تخلف عنه أضاع نفسه وأضاع غيره.

وأنا أكبر في هذه الجامعة بأنها كانت سبَّاقة في تنظيم هذا المؤتمر بشكل سنوي، ولكنها أيضاً تتعامل يومياً مع العديد من الملفات المتعلقة بهذا الجانب الحيوي، ففي مجال أمن المعلومات التقينا على هذه الطاولة، وكذلك في مجال تكريم المبدعين، وفي مجال التعليم الإلكتروني، وفي مجال تنسيب وتنشيط وتفعيل وتدريب أبنائنا للمشاركة في المسابقات العالمية مايكروسوفت وغيرها.

لقد أثبتت الجامعة أن لها نسيجاً مميزاً بتبنيها فلسفة التعليم المفتوح والتعليم الإلكتروني الذي اعتمدته كثير من الجامعات العريقة في العالم منذ زمن طويل في حين ما زال هذا النمط التعليمي لا يحظى بالتقدير الذي يستحقه من الكثيرين، تماماً كما ينظر لموضوع التجارة الإلكترونية، والتداوي عن بعد وغيرهما من الإنجازات التكنولوجية التي ميَّزت عالمنا المعاصر.

بسم الله الرحمن الرحيم

كلمة مستشار الرئيس
لشؤون الاتصالات وتكنولوجيا المعلومات
د. صبري صيدم

بالأمس القريب، –وقبل أيام معدودة– أطلقنا الشبكة العلاجية الإلكترونية الأولى في فلسطين لنربط ثلاثة مستشفيات في الضفة مع مستشفيين في قطاع غزة مع مركز تدريبي في رام الله للتواصل فيما بينها من ناحية، ومع بالعالم الخارجي من ناحية أخرى، وهدفنا من ذلك أن نوفر على المواطن المكلوم مشقة السفر وعناءه وتكلفته المالية الباهظة.

وبالأمس كنا بصحبة د. يونس وزملائنا من الجامعة في إطلاق مؤتمر البلديات الإلكترونية، وقبل ساعة من الآن وقعت اتفاقية بين مجموعة من الشركاء المحليين والأوروبيين لينطلق بموجبها المركز الفلسطيني للإبداع في مجال المال والأعمال. وها نحن الآن على المنصة نحتفي بسنَّة حميدة سنَّتها جامعة القدس المفتوحة من خلال تنظيم هذا اللقاء.

وفي كل ما ذكرت كانت جامعة القدس المفتوحة حاضرة وموجودة بأكاديمييها وطلابها الذين شكلوا العمود الفقري في إبراز هذه النشاطات التي أسلفت ذكرها، بالإضافة إلى أن رئيس الجامعة، موجود على طاولة السياسة، كما هو موجود على طاولة الأكاديمية، وموجود أيضاً على طاولة الإبداع والتميز، وهذه رسالة الشعب الفلسطيني يحزن بالأمس، ويفرح اليوم ويحضر مؤتمراً ويناقش رسالةً، يحضر عرساً، ويذهب إلى بيت عزاء، هذه هي رسالة الشعب الفلسطيني ونسيجه الذي يعتز به ويفخر، ولكن رسالتي الأكبر لك يا د. يونس: كل المبدعين من أبنائك وطلابك وأحبتك، لم يكن

أما في مجال خدمة الطلبة، فقد أصبح طلبة الجامعة اليوم في راحة كاملة في التعامل مع شؤونهم الإدارية والأكاديمية من خلال الجهود التكنولوجية التي تقدمها الجامعة من حيث بناء البرامج الإلكترونية الخاصة بخدمة الطلبة من البوابة الأكاديمية إلى البث الفيديوي التدفقي Video Streeeming وإلى المودل ١ أو المودل ٢ الذي وضع لخدمة الطلبة من حيث التعليم الإلكتروني.

أرحب بكم اليوم أجمل ترحيب وأثمن مشاركتكم معنا في حضور أعمال هذا الملتقى الدراسي، وآمل أن يخرج المؤتمر بتوصيات تسهم في تطوير قدرات الجامعة في هذا القطاع الحيوي، وأتوجه بالشكر الجزيل للأخوة الباحثين وآمل أن تسهم أوراقهم العلمية في تحقيق أهداف المؤتمر، ولا يفوتني أن أتوجه بالشكر لكل من أسهم في عقد هذا المؤتمر إعداداً وتنظيماً ومتابعة.

والسلام عليكم ورحمة الله وبركاته

بسم الله الرحمن الرحيم

كلمة رئيس جامعة القدس المفتوحة
أ.د. يونس عمرو

دأبت جامعة القدس المفتوحة منذ بدايتها على الاهتمام الخاص بالتكنولوجيا والاتصالات، واستثمرت جهوداً طيبة على المستويين: العلمي والأدبي، والأخلاقي والمادي في هذا القطاع، وسخرته إلى حد مميز في سبيل تطوير التعليم الجامعي وتفعيله، فأنشأت المركز المعروف بمركز الاتصالات والمعلومات في الجامعة ICTC، كذلك أنشأت مركز التعليم الإلكتروني المعروف بـ OLC، كذلك استثمرت الجامعة مبالغ مالية لا يستهان بها في هذا القطاع، وغدا هذا الاستثمار –في الحقيقة– منتجاً وله مردود مالي يعود بالنفع على الجامعة، واعتقد أن نائب الرئيس للشؤون المالية يعلم ذلك، وخاصة في مجالات التدريب، إذ إن الجامعة اليوم لها علاقات واتفاقيات مع مؤسسات محلية رسمية وأهلية ومع مؤسسات خارجية في هذا القطاع، وأصبح الاستثمار مثمراً للجامعة.

تقدمت الجامعة إلى حد كبير في موضوع التكنولوجيا والاتصالات بحيث أصبحت تسعى إلى تكريس التعليم المدمج الذي يجمع في نسق واحد بين التعليم المفتوح والتعليم الإلكتروني أو التعليم الافتراضي، وحققت كذلك نتائج طيبة في السوق الإلكتروني، وفي مجالات التدريب، ومجالات أمن المعلومات على مستوى الوطن، وأصبح الباحثون في هذا الميدان من جامعة القدس المفتوحة يشقون طريقهم في المجتمع المحلي في هذا القطاع، وكذلك يشاركون في مؤتمرات محلية وخارجية، وأثبتوا كفاءة عالية بشهادة كل من يزور الجامعة ويطلع على جهودهم.

رسالة رئيس المؤتمر

باسمي وباسم اللجنة المنظمة للمؤتمر، أود أن أشكر الأستاذ الدكتور يونس عمرو (رئيس الجامعة) على دعمه للمؤتمر الدولي لتكنولوجيا المعلومات والاتصالات : التطبيقات والتقنيات. عمل هذا المؤتمر على ايجاد فرصة للمشاركين في المؤتمر على تبادل الأفكار والبحوث في مجال تكنولوجيا المعلومات والاتصالات المختلفة.

نظم هذا المؤتمر من قبل جامعة القدس المفتوحة – فلسطين، وكان يركز على أحدث التقنيات المتعلقة بالمعلومات والاتصالات الرقمية. ويعكس هذا التركيز اهتماماً متزايداً ومتنامياً بكل ما يُستجد من موضوعات أو مجالات تتعلق بتكنولوجيا المعلومات والاتصالات، التي تشكل الموضوع الرئيس للمؤتمر مع ما ينبثق عنه من محاور فرعية تم وضعها بين أيدي الدارسين والباحثين لدراستها ومعالجتها.

نامل ان يكون هذا المؤتمر مثمر، ذو طابع ثقافي وعلمي وممتع. ونحن نتطلع إلى تلقي تعليقاتكم البناءة التي من شأنها أن تساعدنا في التخطيط المستقبلي. لتطوير المؤتمر. يرجى زيارة موقعنا على الانترنت : (http://www.qou.edu/icict2011/index.jsp) للحصول على معلومات عن المؤتمرات التي سيتم عقدها في المستقبل القريب والمجلات البحثية ذات الصلة بالمؤتمر.

وأود أن أغتنم هذه الفرصة لأشكر الفريق المنظم للمؤتمر على العمل المميز للتحضير لهذا المؤتمر. واتقدم بالشكر والامتنان الى اعضاء اللجنة العلمية والمحكمين الذين عملوا على تحكيم الابحاث لاختيار أفضل الابحاث التي تتناسب مع محاور المؤتمر.

وأخيرا ، وبالنيابة عن جامعة القدس المفتوحة المستضيفة والمنظمة للمؤتمر ممثلة بالأستاذ الدكتور يونس عمرو رئيس الجامعة ودوائرها ومراكزها المختلفة وكلية تكنولوجيا المعلومات والعلوم التطبيقية، واللجنة العلمية والراعي المالي للمؤتمر "جوال" ، نشكر جميع المشاركين والباحثين في هذا اللمؤتمر الدولي لتكنولوجيا المعلومات والاتصالات (ICICT'2012)، الذي عقد في جامعة القدس ألمفتوحة رام الله، فلسطين بتاريخ ٢٦، يوليو، ٢٠١٢. ونحن على ثقة بأن هذا المؤتمر قدم المعلومات المفيدة وعمل على تعزيز التعاون الدائم في محاور ومواضيع المؤتمر.

د. يوسف ابوزر

رئيس اللجنة المنظمة والعلمية للمؤتمر

نقدّر ونشكر الراعي والممول الحصري للمؤتمر

جامعة القدس المفتوحة
عمادة البحث العلمي والدراسات العليا

فلسطين/ رام الله - الماصيون

ص.ب. 1804

ت: 02\2984491 - 02\2952508

ف: 02\2984492

البريد الإلكتروني: sprgs@qou.edu

تصميم: مركز تكنولوجيا المعلومات والإتصالات

جامعة القدس المفتوحة

عمادة البحث العلمي والدراسات العليا

المؤتمر الدولي
لتكنولوجيا المعلومات والاتصالات

التطبيقات والتقنيات *(ICICT'2012)*