IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 4 Issue 4

April 2013

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the April 2013 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 4 Issue 4 April 2013
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

E-learning in Higher Educational Institutions in Kuwait: Experiences and Challenges

Mubarak M Alkharang, George Ghinea
Department of Information Systems and Computing
Brunel University
London, UK

***Abstract*—E-learning as an organizational activity started in the developed countries, and as such, the adoption models and experiences in the developed countries are taken as a benchmark in the literature. This paper investigated the barriers that affect or prevent the adoption of e-learning in higher educational institutions in Kuwait as an example of a developing country, and compared them with those found in developed countries. Semi-structured interviews were used to collect the empirical data from academics and managers in higher educational institutions in Kuwait. The research findings showed that the main barriers in Kuwait were lack of management awareness and support, technological barriers, and language barriers. From those, two barriers were specific to Kuwait (lack of management awareness and language barriers) when compared with developed countries. Recommendations for decision makers and suggestions for further research are also considered in this study.**

Keywords–e-learning; higher education; adoption; Kuwait; developed countries; e-learning barriers.

I. INTRODUCTION

E-learning has emerged as a necessity to meet the challenges posed by the development of information technology and its potential for greater access to knowledge [1]. E-learning was first introduced in developed countries; thus, the adoption and utilization models developed there have been taken as benchmarks worldwide. Essentially, the influential factors and barriers to the adoption of e-learning within different societies and region may or may not be the same as for those identified in developed regions, with varying degrees of intensity or importance [2]. Accordingly, the models available for adoption may not be applied across all steps and phases when utilized by different societies and countries. As such, influential factors and barriers to e-learning may vary between cases.

In regard to educational establishments across the globe, e-learning is becoming more widely adopted. As with many different countries, the adoption of e-learning in the context of higher educational institutions has become the subject of much research and examination. Importantly, regardless of the high standards of living within the country, Kuwait is falling behind other countries because of its relatively poor innovation and productivity capabilities [3]. With this noted, it is essential that organizations and the government work together in order to update and upgrade the skills of their subjects, whether employees, customers or students, and to further deliver on-going learning and training where e-learning is still to play a key role [4].

The aim of this research is to investigate and identify factors that will mostly influence the adoption of e-learning in Kuwait as an example of a developing country, and compare them to those found in developed countries. In order to achieve this aim, our research was conducted in two phases. The first was based on reviewing and examining the existing literature and studies on e-learning, highlighting the factors that influence the adoption of e-learning. The second phase comprised empirical data collection, where semi-structured interviews were conducted in Kuwait. The results of this study will help decision makers to gain a better understanding of the factors that determine and influence the adoption of e-learning in higher educational institutions in Kuwait.

The structure of this paper is as follows; the next section introduces e-learning in terms of definition, history, and advantages. Then, the main factors and barriers to e-learning adoption in developed countries are highlighted. After that, the empirical background to this study is presented and followed by the research methodology. The following section discusses and compares the research findings. Finally, the paper concludes by discussing the implications of the study's findings, and identifying future research directions.

II. INTRODUCTION TO E-LEARNING

A. Definition of e-Learning

The term e-Learning devotes to electronic learning, which was defined by the NCSA e-Learning group as "the acquisition and use of knowledge distributed and facilitated primarily by electronic means. This form of learning currently depends on networks and computers but will likely evolve into systems consisting of a variety of channels (e.g., wireless, satellite), and technologies (e.g., cellular phones, PDAs), as they are developed and adopted. E-learning may incorporate synchronous or asynchronous access and may be distributed geographically with varied limits of time" [5]. While Koohang & Harman [6] defined e-learning as:

> *"The delivery of education (all activities relevant to instructing, teaching, and learning) through various electronic media. The electronic medium could be the Internet, intranets, extranets, satellite TV, video/audio tape, and/or CD ROM."*

B. History of e-Learning

Since the 1960s, educators and trainers at all levels of education, business, training and military made use of technology and computers in different ways to support and enhance learning [7]. Accordingly, in the early 1960s, Don Bitzer at the University of Illinois created a timeshared computer system called PLATO that was concerned with literacy programs. PLATO allowed students and teachers to use graphics terminals and TUTOR, an educational programming language, to communicate and interact with other users by means of electronic notes, thus being the forerunner of today's conferencing systems [7].

Nowadays, e-learning is evolving with the World Wide Web as a whole and it's changing to a degree significant enough to warrant a new name, namely e-learning 2.0. The term e-Learning 2.0 is used to refer to new ways of thinking about e-learning inspired by the emergence of Web 2.0. From an e-Learning 2.0 perspective, e-learning will concentrate on social learning and the use of social software such as blogs, wikis, podcasts and virtual worlds.

C. Benefits and Drawbacks of e-Learning

The significant shift towards e-learning is clearly motivated by the various advantages it offers. Despite the fact that e-learning has received much commendation, it remains that human instructors will never be replaced completely by computer systems [8]. Nevertheless, establishing the benefits achievable through e-learning is important. It is recognized that some of the key benefits include the reduction of overall cost (instructors' salaries, travel costs, and meeting room rentals), as well as access to quality education, the provision of convenience and flexibility, a reduced environmental impact through lower paper use and energy consumption, and higher retention [9, 10].

Although the benefits and advantages of e-learning are obvious, it could be argued that it still has some weaknesses. For example, since users are not bound by time, the course is available 27/7 and does not require physical attendance which could reduce the social and cultural interaction. The learners may also feel isolated and unsupported while learning since the instructors and instructions are not always available. They may become bored with no interaction. Technology issues required for e-learning could also become potential problems for the learning process. The learners need to have access to resources such as computers, internet, and software. They also need to have computer skills with programs such as word processing, Internet browsers, and e-mail [11, 12].

III. BARRIERS TO E-LEARNING ADOPTION

Reviewing the literature on e-learning practices shows common agreement on the importance of information and communication technology (ICT) in today's learning environment [13]. Most organizations have understood that e-learning has to be integrated as part of daily tasks of students and employees (academics and managers), not to be seen as a separate tool or technique for learning and training. Therefore, e-learning has become a strategic advantage that participates in the realization of the organizational strategic plan [14].

Figure 1 shows the model of e-learning development found in developed countries and is adopted from the reviewed literature [2]. E-learning initiatives come about as a result of environmental trends which are made up of the same factors that are the cause and the consequence of the revolutionized technology, i.e. rapid change of technology and rapid change to learning environment. The outcome of these trends has been an increasing amount of pressure on educational organizations to use IT to improve their capacity to respond to learning needs. From this pressure, the decision to implement e-learning emerges.

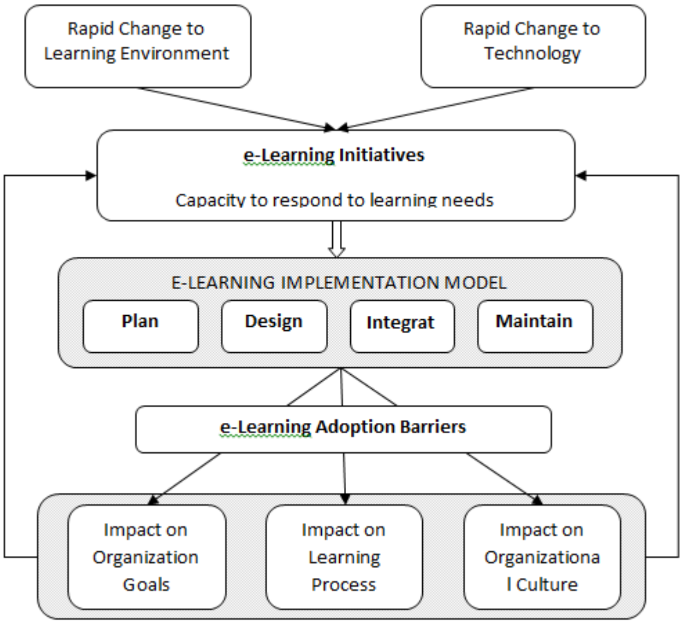

Rapid Change to Learning Environment
Rapid Change to Technology
e-Learning Initiatives
Capacity to respond to learning needs
E-LEARNING IMPLEMENTATION MODEL
Plan
Design
Integrat
Maintain
e-Learning Adoption Barriers
Impact on Organization Goals
Impact on Learning Process
Impact on Organizationa l Culture

Fig.1. Development Model of e-Learning

The barriers to e-learning adoption can be found between the decision to implement e-learning and the impacts on the organization. The e-learning implementation model follows the usual implementation model of any type of information system, and it is made up of four stages: planning, designing, integrating, and improving [15]. The 'integration' stage marks the occasion when the system is put to work in the organization (i.e. the organizational implementation). The last stage of 'improvement' can only be carried out after the impacts on the organization have been evaluated, i.e. the impact on organization goals, the impact on learning process, and the impact on organizational culture.

Technology critics consistently argue for a balanced review of any technology, but the threats, challenges, and losses brought by technology are typically less discussed. While focusing on barriers might be construed negatively, it is not intended to dissuade organizations or individuals from using learning technologies. Rather, the aim is to increase awareness and understanding in regard to the overall nature of issues experienced by e-learners. Such an examination is critical considering the heavy investments, promises, and exponential growth associated with e-learning.

Thus, according to Bernárdez [16], the barriers to the implementation and adoption of e-learning can be related to personal issues, technical issues, or organizational issues. The most commonly cited personal barriers are time management problems, language problems, attitude towards e-learning and learning styles or preferences where learners might prefer passive or active learning. The technical barriers include infrastructure building and upgrading, maintaining connectivity and bandwidth, accessibility and usability, and lack of technical support. The organizational barriers include lack of e-learning awareness, lack of management support and commitment, lack of strategic planning and direction, lack of time available for learning and training, lack of appropriate content and assessments, and lack of incentives and credibility [17, 18].

After reviewing the literature and highlighting the main problem areas, important factors and key barriers were grouped into four main categories: cost, time, technology, and attitude. Importantly, cost is commonly highlighted as being one of the most notable of barriers facing e-learning. Technology is fundamental within e-learning, and it is also expensive and unpredictable, which makes the initial costs of implementation and on-going costs of maintenance very high [19, 20].

Moreover, time here refers to the amount of time required to establish and maintain e-learning within organizations. It also refers to the amount of time made available by both organizations and learners for e-learning. Considering external interruptions and distractions to the learners and maintaining the appropriate concentration for e-learning, the time factor is a significant barrier to e-learning. In fact, it occupies an important rank among the top barriers to adopting e-learning in organizations [20, 21].

Likewise, technology is critical in adopting e-learning. It requires adjustments from both, learners and organizations. For organizations to effectively implement e-learning, they need to ensure that they have the appropriate capacity to run e-learning systems and that serious consideration is given to hardware compatibilities and capabilities. Inadequate software, limited bandwidth and connectivity, and system breakdowns are other problems [22]. Technical support is also a significant issue, especially in cases where the suppliers do not provide this service. In such a scenario, the users may become reluctant to use e-learning. Horton [23] for instance claims that many e-learning courses are dead on arrival due to an inability to install the right software, establish connectivity, or provide support.

Furthermore, the attitude towards e-learning is important. To achieve the promises of e-learning, users need to embrace it and management needs to provide the necessary support. Essentially, if they feel that e-learning creates more problems than it solves or if they simply do not know how to use it, all involved will not be comfortable with its use, and will therefore demonstrate signs of resistance, thus rendering the e-learning system valueless [20, 22].

IV. ADOPTION OF E-LEARNING IN KUWAIT

The ever-growing use and adoption of information and communication technologies by the Kuwaiti government departments and organizations have helped build an IT infrastructure capable of adopting new technologies such as e-commerce, e-government, and e-learning [24]. The term e-learning is relatively new in Kuwait, and only a limited number of local suppliers offer e-learning systems, currently implemented in some of petroleum and financial companies. Nevertheless, an increasing number of organizations are responding to the challenge of e-learning and are moving to adopt it, yet are finding significant barriers hampering their efforts [3, 25].

Importantly, however, there is a limited number of studies on the field of e-learning implementation and adoption in Kuwait, and very few researches have been carried out on the barriers encountered by organizations and higher educational institutions using this relatively new learning method [2, 25]. It should be acknowledged that, in the country, technological use is expanding, and the use of the Internet, more specifically, is garnering much attention. According to the Arab Human Development Report [26], the number of Internet users in 1990 was 0%, while in 2003 the number was 23% and in 2009 this had increased to 37%. Hence, while considering the international movements towards an information-based society and underlining e-learning as a means to improve the learning and training of organizations, the evolution of technology and Internet in Kuwait and the barriers and recommendations that will be addressed in this study should be taken into consideration when adopting any e-learning system in Kuwait.

As stated before, e-learning as an organizational activity started in the developed countries, and as such, the implementation models developed in the developed countries are taken as a benchmark. Furthermore, the factors and barriers that influence the adoption of e-learning in different regions and societies may or may not be the same as those found in the developed countries with varying degrees of intensity or importance. Hence, those available implementation models may not necessarily be followed in all stages and steps when used by different countries and societies. Accordingly, the implementation barriers and the influential factors may differ from one case to another.

V. RESEARCH METHODOLOGY

A qualitative approach was used for this study in order to explore and study emerging phenomena within their context. Based on Denzin and Lincoln’s [27] recommendations, the data collection was carried out through direct contact with the main higher educational institutions in Kuwait, where there are five universities and six colleges. Of these, there is only one public university (Kuwait University, KU) and one public college (Public Authority of Applied Education and Training, PAAET). Six higher educational institutions were chosen, representative of the biggest numbers of students and staff in Kuwait.

The study was limited to higher educational institutions’ academics and managers in Kuwait, where fifteen members of the chosen organizations were contacted. The sample was chosen for convenience and practical reasons since knowing their opinions and perceptions will help to improve the services provided by this technology. Further, they were chosen because universities’ academics and managers are amongst those whose attitudes and supports will influence the adoption of e-learning in their organizations [3].

The data collection was based on semi-structured interviews [28-30]. Questions covered in the interview guide were laid out in three sections. The first section targeted general and historical background information on e-learning in the organization. The second section sought to identify the barriers and challenges faced in implementing and adopting e-learning in the organization, focusing on identifying the most significant barriers by both management and users. The third section was about evaluating the e-learning experience in the organization.

The interview guide was reviewed and evaluated by three e-learning practitioners and researchers. Based on their pilot evaluations and recommendations, the questions were revised and modified [31]. The interviews were conducted over a period of 6 weeks. Most of the interviews lasted between 60 and 90 minutes. Each interview was tape recorded and transcribed. These were given back to each participant to check any differences that may have arisen and to eliminate any bias [32].

In this study, the data were analyzed using thematic analysis [33], where the qualitative information were encoded in order to identify specific themes; that is, whether some sort of patterns are identified within the information that may have some relevance to the area of research [33, 34]. Thematic analysis steps suggested by Braun and Clarke [35] were followed. These steps start by reading and familiarizing with the data, generating initial codes by organizing the data, searching for themes by re-reading and reviewing the data, and defining and naming these themes.

VI. RESEARCH FINDINGS AND DISCUSSION

The research findings offer insights into the main and influential factors that influence the adoption of e-learning in higher educational institutions in Kuwait. After summarizing the data collected and highlighting the main points, common themes were regrouped and key points and problem areas were divided into three main categories. These categories are management awareness and support, technology, and language barriers.

A. Management Awareness and Support

The main part of the study was to identify the key factors influencing the organizations surveyed from building an environment supportive of e-learning. The vast majority of those questioned on the limitations of e-learning (12 out of 15 respondents) stated lack of management awareness and support as the main barrier. In most cases, the strategy of the management in the organization was not in line with the intention to build an e-learning culture. E-learning was seen by the management as a waste of time process and an ineffective option for learning. One of the IT specialist said "Of course, a supportive management is a key factor for the acceptance of any new project including e-learning. However, the management will not support e-learning unless they are aware of the benefits it offers, and unfortunately our management is unaware of the benefits and strategic advantages of e-learning". This was more obvious in the replies of interviewees working in public higher educational institutions in Kuwait. In such cases, the top management are more concerned with their own image and profit, rather than the organization's image. However, in the private educational institutions, the management is more concerned with a return on investment and therefore adopting e-learning has a higher priority than in public organizations.

Since the management was the source of resistance, the lower level employees did not sincerely buy into the e-learning projects. There was a "*lack of understanding about e-learning"*, as one of the respondents mentioned in describing the organization environment. As a result, even when e-learning did deliver benefits, they were hampered by the inter-group conflict in the organizations. Other interviewees stated that the management lacked the awareness of the strategic benefits of e-learning. Such a lack of awareness was felt through the absence of clear training and learning policies aimed at developing the knowledge and skills of their staff. Some interviewees mentioned that some managers and academics were computer illiterate; thus, they were afraid of the new technology and more comfortable with the traditional methods. One of those interviews said *"How would you convince those old people to use e-learning while they don't know how to use computers?"*. Interestingly, an academic stated that the content development in the e-learning modules was very poor and there was a limited involvement in the contents development process. As a consequence, many academics did not feel motivated to use the e-learning system and showed high levels of resistance and reluctance.

Nevertheless, the key lesson which has been derived from this factor is that the problem is not one of structure but of processes. The difficulty consists in knowing the management processes that lead to a successful adoption of e-learning. The management in the surveyed educational organizations has failed to understand the strategic advantages of using e-learning as a means to improve the learning process.

B. Technology Barriers

Technology problems came high in the list of barriers in Kuwait where they were mentioned by 10 out of 15 participants. Bandwidth and internet speed limitations were seen as significant barriers to starting and adopting e-learning in the educational organizations. Some interactive tools and multimedia simulations take far too long for the user to access and use. This was apparent in organizations that did not have appropriate infrastructure that support the e-learning system. Furthermore, technology standards were seen by IT specialists in the organizations as an important requirement for e-learning success. One said that *"we need to standardize the procedures, formats and systems within the organization"*. Those standards act as the base to use physical and intellectual IT assets. In addition, some academics said that the absence of technical support would be a triggering factor for ending the e-learning project since most of the users are not familiar with e-learning technologies and procedures.

The findings also revealed that more than half of participants (8 out of 15) were worried about security and confidentiality issues. Security and confidentiality concerns are seen mostly by academics as one of the most important issues due to the sensitivity of information being transferred online such as assessments and grades. A system administrator said *"Security aspects and data confidentiality are very important to accept and use e-learning by academics. However, they have less effect on the students' perceptions"*. Security issues include computer and network security, privacy and confidentiality of data. Underrating the importance of this factor could cause unauthorized access to sensitive information and loss of users' trust, which might hinder the adoption of e-learning. Surprisingly, one member of the top management was concerned with system integration where local systems are linked together and contain all different functions which would provide a full and real one stop shop. It is common for different departments to have different software and hardware that may not work together which may lead to e-learning implementation and adoption difficulties.

The technological problems mentioned by the interviewees were critical for the adoption of e-learning in Kuwait. Regardless of the fact that the necessary resources and equipment (personal computers in particular) for using e-learning were made available in most of the educational organizations surveyed, all interviewees mentioned that there was plenty of room for improvement and the intensity of barriers was strong enough to wear away the positive effects obtained from e-learning. This is mostly due to the lack of appropriate implementation of the e-learning implementation model mentioned earlier (Figure 1). The implementation model of e-learning in Kuwait was not made up of the four usual stages but only one: integration, while there was not much concern about planning, designing or evaluating the e-learning investments.

C. *Language Barriers*

Language barriers were found to be significant barriers, having been mentioned by 9 out of the 15 interviewees. Most of the e-learning contents used in the organizations were developed in English, and many of those organizations had a large number of employees who did not master the English language. Those who did not master the English language either conducted their education in non-English countries or their fields are not English specialty, hence they were reluctant to use e-learning. Language barriers were also mentioned by academics who feel students in higher educational institutions in Kuwait will feel uncomfortable when using e-learning courses that were developed in English. Those students have normally undertaken all their previous education in Arabic, speaking English as a second language and have varying levels of English proficiency. One interviewee mentioned that some departments in the organization were not English literate; and thus, they were afraid of the new system that does not support their language. He said "departments such as Law and Arabic Literature provide their teaching and course contents in Arabic, hence the academics and students there will be reluctant to use the e-learning system if not customized and translated to Arabic".

Meanwhile, the cost for developing Arabic contents and courses was very expensive and logistically complicated for many of these organizations. The Arab Human Development Report [26] urged the governments and policy makers in the Arab states to encourage and reward professionals and entrepreneurs to develop content in Arabic that incorporated different aspects of the culture and tradition and publish it on the Internet. It seemed that progress was still slow in content development and the organization environment still relied on contents and courses developed in English for e-learning. The fact that language is recognized as a significant barrier in Kuwait is a reflection of the ready-to-wear approach that the participated organizations have followed with little consideration to content appropriateness or culture.

VII. BARRIERS COMPARISON BETWEEN KUWAIT AND DEVELOPED COUNTRIES

A comparison (using a simple 3-point scale, irrelevant, relevant, or important) of barriers between the Kuwaiti and developed countries experiences is shown in Table 1. The comparison was established by ranking the barriers according to their degree of importance.

TABLE I. COMPARISON OF BARRIERS BETWEEN KUWAIT AND DEVELOPED COUNTRIES

Barriers	Developed Countries	Kuwait
Cost	Important	Irrelevant
Time	Important	Relevant
Technology	Important	Important
Attitude	Important	Relevant
Management Awareness and Support	Relevant	Important
Language	Irrelevant	Important

From the literature [19-22], the order of priorities of barriers found in developed countries was cost, time, technology, attitude, management awareness and support, and language. In comparison, the order of priorities found in Kuwait in this study was management support, language, technology, attitude, time, and cost. Looking at the two rankings, only technology shares the same degree of importance. However, in the remaining barriers, Kuwait varies from developed countries which support our view in that the barriers in different regions and societies may or may not be the same as those found in the developed countries with varying degrees of intensity or importance. In the developed countries, cost, time, and attitude barriers ranked highly as important. In Kuwait, time and attitude appeared as relevant, and cost was considered irrelevant. On the other hand, management support and language barriers were rated as highly as important in Kuwait, whereas in developed countries, management support barriers are considered relevant and language barriers considered irrelevant.

In Figure 1, the decision to implement e-learning in developed countries is followed by an implementation model which consists of the usual four stages: planning, designing, integrating and improving. From the research findings, the implementation model of e-learning in Kuwait did not embrace the four usual stages but only one: integration. Unfortunately, there was not much concern in the participating organizations about planning, designing or evaluating the e-learning investments.

VIII. Conclusion

The aim of this research is to investigate and study the factors that influence the adoption of e-learning in higher educational institutions in Kuwait, and help to reduce the resistance towards using e-learning. Therefore, we have to find and study all the factors that influence the success of e-learning in order to successfully reach the adoption of the e-learning. In this study we have introduced and analyzed some factors that influence the acceptance and adoption of e-learning in developed countries in general, and specifically in Kuwait as an example of a developing country. The importance and intensity of barriers to e-learning adoption found in developed countries were found to be different from those found in Kuwait. Furthermore, the implementation model of e-learning established in developed countries was not followed in all stages when implemented in Kuwait. The research findings confirm the research assumptions in that the factors and barriers that influence the adoption of e-learning in different regions and societies may not be the same as those found in the developed countries with varying degrees of importance. Hence, those available implementation models may not necessarily be followed in all stages and steps when used by different countries and societies.

Limitations to this study revolved around the lack of up-to-date data on the state of using e-learning in higher education institutions in Kuwait. There were no formal statistics on e-learning projects and plans for those specific institutions. Lastly, the hectic schedule of some of the participants made it difficult to complete the interviews without interruptions. In spite of those limitations, it is believed this study would offer firm indications about the current state of e-learning, and provides a contribution to the growing literature on e-learning in Kuwait.E-learning adoption in developing countries is a huge project, with many criteria that may not be necessary the same as those found in the developed countries. In Kuwait, the impact of e-learning so far has failed to deliver benefits, and improvements have rarely matched expectations. The lack of awareness of the potential benefits to be accrued from creating an e-learning culture in which technology is a facilitator does create some confusion. However, there were indications and hopes that it is only a matter of time before e-learning is fully appreciated and adopted in Kuwait.

References

[1] Bottino, R.M., *The evolution of ICT-based learning environments: which perspectives for the school of the future?* British Journal of Educational Technology, 2004. 35(5): p. 553-567.

[2] Ali, G.E. and R. Magalhaes, *Barriers to implementing e-learning: a Kuwaiti case study.* International journal of training and development, 2008. 12(1): p. 36-53.

[3] Aldhafeeri, F., M. Almulla, and B. Alraqas, *TEACHERS EXPECTATIONS OF THE IMPACT OF E-LEARNING ON KUWAITS PUBLIC EDUCATION SYSTEM.* Social Behavior and Personality: an international journal, 2006. 34(6): p. 711-728.

[4] Al-Kazemi, A.A. and A.J. Ali, *Managerial problems in Kuwait.* Journal of Management Development, 2002. 21(5): p. 366-375.

[5] Meyen, E.L., et al., *e-Learning: A programmatic research construct for the future.* Journal of Special Education Technology, 2002. 17(3): p. 37-46.

[6] Koohang, A. and K. Harman, *Open source: A metaphor for e-learning.* Informing Science: International Journal of an Emerging Transdiscipline, 2005. 8: p. 75-86.

[7] Nicholson, P., *A History of E-Learning*, in *Computers and Education*, B. Fernández-Manjón, et al., Editors. 2007, Springer Netherlands. p. 1-11.

[8] Mouzakitis, G.S., *E-Learning: The six important "Wh…?".* Procedia - Social and Behavioral Sciences, 2009. 1(1): p. 2595-2599.

[9] Gill, M., *``E-learning technology and strategy for organisations''.* The Business of E-learning: Bringing your Organization in the Knowledge E-conomy, University of Technology, Sydney, 2000.

[10] Roy, R., S. Potter, and K. Yarrow, *Designing low carbon higher education systems: Environmental impacts of campus and distance learning systems.* International journal of sustainability in higher education, 2008. 9(2): p. 116-130.

[11] Welsh, E.T., et al., *E-learning: emerging uses, empirical results and future directions.* International Journal of Training and Development, 2003. 7(4): p. 245-258.

[12] Collins, C., D. Buhalis, and M. Peters, *Enhancing SMTEs' business performance through the Internet and e-learning platforms.* Education+ Training, 2003. 45(8/9): p. 483-494.

[13] Lytras, M.D., A. Pouloudi, and A. Poulymenakou, *Knowledge management convergence–expanding learning frontiers.* Journal of knowledge management, 2002. 6(1): p. 40-51.

[14] Magalhães, R., *Organizational knowledge and technology: an action-oriented perspective on organization and information systems*. 2004: Edward Elgar Publishing.

[15] Alter, S., *Which life cycle---Work system, information system, or software.* Communications of the AIS, 2001. 7(17): p. 1-52.

[16] Bernárdez, M., *From E-Training to E-Performance: Putting Online Learning To Work.* Educational Technology, 2003. 43(1): p. 6-11.

[17] Al-Shery, A., et al. *The Motivations For Change Towards E-Government Adoption: Case Studies From Saudi Arabia*. 2006.

[18] Mungania, P., *The seven e-learning barriers facing employees.* Retrieved November, 2003. 18: p. 2004.

[19] Murray, D. *E-learning for the workplace: Creating Canada's lifelong learners*. 2001: Conference Board of Canada.

[20] Simmons, D.E., *The forum report: E-learning adoption rates and barriers.* The ASTD e-learning handbook, 2002: p. 19-23.

[21] Baldwin-Evans, K., *Employees and e-learning: what do the end-users think?* Industrial and Commercial Training, 2004. 36(7): p. 269-274.

[22] Netteland, G., B. Wasson, and A.I. Mørch, *E-learning in a large organization: A study of the critical role of information sharing.* Journal of Workplace Learning, 2007. 19(6): p. 392-411.

[23] Horton, W., *e-Learning by Design*. 2011: Wiley.

[24] Al-Fadhli, S., *Instructor Perceptions of E-learning in an Arab Country: Kuwait University as a case study.* E-Learning and Digital Media, 2009. 6(2): p. 221-229.

[25] Al-Fadhli, S., *Factors Influencing the acceptance of distance-learning.* International Journal of Instructional Media, 2011.

[26] UNDP, *Arab Human Development Report-UNDP (2009), Challenges to Human Security in the Arab Countries*. 2009.

[27] Denzin, N.K. and Y.S. Lincoln, *The discipline and practice of qualitative research.* Handbook of qualitative research, 2000. 2: p. 1-28.

[28] Myers, M.D. and D. Avison, *Qualitative Research in Information Systems~ autofilled~*. 2002.

[29] Patton, M.Q., *Qualitative research*. 2005: Wiley Online Library.

[30] McMillan, J.H. and S. Schumacher, *Research in education*. 2009: Pearson Education.

[31] Presser, S., et al., *Methods for testing and evaluating survey questions.* Public opinion quarterly, 2004. 68(1): p. 109-130.

[32] Irani, Z., et al., *Evaluating e-government: learning from the experiences of two UK local authorities.* Information Systems Journal, 2005. 15(1): p. 61-82.

[33] Bradley, E.H., L.A. Curry, and K.J. Devers, *Qualitative data analysis for health services research: developing taxonomy, themes, and theory.* Health services research, 2007. 42(4): p. 1758-1772.

[34] Hsieh, H.-F. and S.E. Shannon, *Three approaches to qualitative content analysis.* Qualitative health research, 2005. 15(9): p. 1277-1288.

[35] Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative research in psychology, 2006. 3(2): p. 77-101.

A Cost-Efficient and Reliable Resource Allocation Model Based on Cellular Automaton Entropy for Cloud Project Scheduling

Huankai Chen
Future Computing Group
University of Kent
Canterbury, UK

Frank Wang
Future Computing Group
University of Kent
Canterbury, UK

Na Helian
University of Hertfordshire
Hatfield
Hertfordshire, UK

***Abstract*—Resource allocation optimization is a typical cloud project scheduling problem: a problem that limits a cloud system's ability to execute and deliver a project as originally planned. The entropy, as a measure of the degree of disorder in a system, is an indicator of a system's tendency to progress out of order and into a chaotic condition, and it can thus serve to measure a cloud system's reliability for project scheduling. In this paper, cellular automaton is used for modeling the complex cloud project scheduling system. Additionally, a method is presented to analysis the reliability of cloud scheduling system by measuring the average resource entropy (ARE). Furthermore, a new cost-efficient and reliable resource allocation (CERRA) model is proposed based on cellular automaton entropy to aid decision maker for planning projects on the cloud. At last, the proposed model is designed using Matlab toolbox and simulated with three basic cloud scheduling algorithm, First Come First Served Algorithm (FCFS), Min-Min Algorithm and Max-Min Algorithm. The simulation results show that the proposed model can lead to achieve a cost-efficient and reliable resource allocation strategy for running projects on the cloud environment.**

Keywords—Resource Allocation; Cloud Project Scheduling; Entropy; Cellular Automaton; Cost-efficiency; Reliability; Complex System; Local Activity; Global Order; Disorder

I. INTRODUCTION

In recent years, Cloud computing is emerging as a new paradigm of large-scale distributed computing, which rent computing resources on-demand, bill on a pay-as-you-go basis, and multiplex many users on the same physical infrastructure. These cloud computing environments provide an illusion of infinite computing resources to cloud users so that they can increase or decrease their resource consumption rate according to the demands. At the same time, resources allocation problem under the cloud environment poses a number of challenges.

Researchers who construct resource allocation strategies for scheduling must cope with the world's natural tendency to disorder. In cloud computing, projects are scheduled on a set of cloud resources that are local active (in the sense that each resource was determined to be assigned tasks based on its own state and the state of the environment and its productivity are affected by the amount of tasks that assigned to it), and corporately structured. We want resource local activity to yield coherent global schedule system order. However, widespread experience warns us that modelling and optimizing systems that exhibit both local activity and global order are not easy. The experience that anything that can go wrong will go wrong and at the worst possible moment is summarized informally as "Murphy's Law" [13]. Scheduling systems are not immune to Murphy. In cloud project scheduling system, after an enough power strikes one of the resources, which leads to its productivity reduced or collapsed, the whole system collapsed. In the real world scenario, such resource productivity reduced or collapsed may cause by hardware/software failures, resources CPU overload, resource over- or under-provisioning, or application misbehaviours. Thus, the system is failed to execute and deliver a project as originally scheduled.

At the root of the ubiquity of disordering tendencies is the Second Law of Thermodynamics, "Energy spontaneously tends to flow only from being concentrated in one place to becoming diffused or dispersed and spread out" [14]. In cloud scheduling system, adding resources to a system may overcome the Second Law "spontaneous tendency" and lead to increasing the system's order. However, the way to decide the numbers of resource allocated to the project is critical. Especially when resources are local active, which is the origin of complexity [12], the scheduling system become more complex under cloud environment. In most case, an increase in the number of assigned resources positively impacts the system's efficiency and reliability. However, there is a limit on the number of assigned resources beyond which any increase may have the opposite effect. Allocate resources beyond this limit may lead to disorder/chaotic condition and a disproportional return on investment in terms of local resource productivity, global system efficiency and reliability.

In the literature, a lot of scheduling algorithms were proposed in the past. Braun et al [6] have studied the relative performance of eleven heuristic algorithms for task scheduling such as First Come First Served (FCFS), Min-Min, Max-Min, Genetic Algorithm (GA), etc. They have also provided a simulation basis for researchers to test the algorithms. A family of 14 scheduling heuristics for concurrently executing BoTs in cloud environments are proposed recently [7]. Most of the past works are mainly aim to shorten project's completion time and enhance the system throughput, which are the focus in improving scheduling algorithm itself. Some theoretical scheduling papers address the reliability problem of scheduling system by analyzing the entropy produced by scheduling algorithm [3] or resource [1]. However, most of those methods treat scheduling problem as a linear programming problem. We argue that such linear programming technique is not suitable for modelling the complex scheduling system which is dynamic and nonlinear. We are not aware of a method that combine the theoretical analysis of scheduling system with using nonlinear modelling, aiming to achieve both cost-efficiency and reliability of resource allocation strategies, and quantitative measuring the relation between local active resource and global system performance all together.

Both efficiency and reliability are the most important factors for planning a project. The reduced efficiency and reliability of the global system is a direct consequence of the disorder caused as a result of the local active resources and the difficulty in managing these resources. Thus, the resulting resource allocation problem is also an entropy-optimization problem: how many resources should be allocated to a project in order to minimize average resource entropy, subject to limited cost budget within the examined time-period. The fundamental claim of this paper is to solve the above cloud resource allocation problem based on Entropy Theory.

Scheduling is an NP-complete problem, the complexity of which increase substantially in the cloud environment. For such class of problems, in order to achieve the optimal solution an effective method for modelling complex system is also needed. A cellular Automata (CA) is a mathematical model for a complex system which evolves in discrete steps [9]. It is suitable for modelling cloud scheduling system which can be described as a massive collection of resources that interact locally with each other [4]. In this paper, we represent the cloud scheduling system behaviour as a cellular automaton, specifically as a one-dimension cellular automata network.

Following the short introduction on the problem, we will begin discussion by recalling in section 2 the detail problem definition and assumptions of resource allocation for cloud project scheduling. In section 3, the paper presents the general concepts of entropy. Cellular Automation will be applied for modelling the complex scheduling system and a resource allocation model based on cellular automaton entropy will be introduced in section 4 and section 5. We will then describe the experiment and present our simulation results in section 6. Section 7 and Section 8 contains some conclusions and possible future research direction.

II. Problem Definition

Due to the NP-completeness nature of a scheduling problem, the developed approaches try to find optimal resource allocation solution with considering both cost-efficiency and reliability in the cloud environment. In this paper, the proposed model has been developed under a set of assumptions:

- A project consists of a collection of tasks that have no dependency among each other. Each task requires amounts of computing demand that are known before the task is submitted for execution, or at the time it is submitted.
- Project needs to be completed within deadline and cost budget
- A collection of numbers of cloud resources is rented for running the project. Resources provide amounts of computing capacity. In this paper, Computing capacities were expressed in EC2 compute units (ECU) [15], which for experimental purposes were defined as 1 EC2 compute unit = 1,000,000 million of instructions per second. Hourly cost rates for one ECU were expressed in USD and were based on the EC2 pricing mode [15].
- Selections of one or more scheduling strategies are available for planning the project on the cloud.

In static heuristics, the computing demand for each task is known a priori to execution and measured by ECU. Thus, the expected execution time for a task running on a resource can be calculated by dividing task computing demand by resource computing capacity.

The main aim of scheduling strategies is to minimize a project's completion time and cost with renting a number of resources within deadline. In such scheduling system, the resource allocation problem can be defined as follows:

Let Task set $T = t_1, t_2, t_3 \ldots t_n$ be the collection of tasks in a project that submitted to execute on the cloud. Each task requires amounts of computing demand $cd_1, cd_2, cd_3 \ldots cd_n$, which is measured by ECU.

Let Resources set $R = r_1, r_2, r_3 \ldots r_m$ be the set of resources that are rented for scheduling the tasks. Each resource has its computing capacity which is also measured by ECU, $cc_1, cc_2, cc_3 \ldots cc_{n.}$

Resources are defined as different types according to their computing capacity [15], Resources Type set $RT = rt_1, rt_2, rt_3 \ldots rt_k$. The Resource Cost Price Rates for different type are $cp_1, cp_2, cp_3 \ldots cp_k$.

The project's completion time, Makespan, can be calculated as follows:

$$Makespan = max(CT_{ij}) \quad (1)$$

$$CT_{ij} = RT_j + ET_{ij}, where\ 1 < i < n, 1 < j < m \quad (2)$$

Where CT_{ij} refers to completion time of task i executing on resource j , ET_{ij} refers to the expected execution time of task i on resource j, and RT_j refers to the ready time of a resource j after completing the previously assigned tasks.

The model we proposed is developed to aid decision makers to solve the following problems:

- How many and what type of resources should we rent?
- How should we schedule the tasks on the rented resources?

So we can achieve a cost-efficient and reliable resource allocation strategy for running the project on the cloud within deadline and cost budget.

III. Entropy Theory

Entropy is an important statistical quantity which measures the disorder degree and the amount of wasted energy in the transformation from one state to another in a system [14]. Although the concept of entropy was originally a thermodynamic construct, it has been adapted in other fields of study, including information theory, production planning, resource management, computer modelling and simulation [1] [2] [3] [5] [11]. We will use this measure to quantify the reliability degree associated with the scheduling system under different resource allocation strategies. First, we introduce this measure in a general content.

Given a dynamic system X of finite mutually exclusive state variable set $S = s_1, s_2, s_3 \ldots s_n$ with probabilities $p_1, p_2, p_3 \ldots p_n$ respectively, entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \times log\, p_i \qquad (3)$$

For any two mutually independent dynamic systems A and B with n and m states respectively, the probability π_{ij} of the of the simultaneous occurrence of the states A_i and B_j is $p_i q_j$ where $p_i$is the probability of state i occurring in system A, q_j is the probability of state j occurring in system B, where $1 \leq i \leq n$ and $1 \leq j \leq m$. Let the sets of states $A_i B_j$ represent another finite system designated by AB. It is easy to see that:

$$H(AB) = H(A) + H(B) \qquad (4)$$

Where $H(AB)$, $H(A)$ and $H(B)$ are the corresponding entropies of systems AB, A and B.

This expression can be easily extended for an arbitrary number of mutually independent finite systems. For a system M consisting of s mutually independent sub-systems $N_1, N_2, N_3 \ldots N_k$, the entropy is given by:

$$H(M) = \sum_{i=1}^{k} H(N_i) \qquad (5)$$

And the average sub-system Entropy [11] is easily obtained by:

$$\overline{H} = \frac{H(M)}{k} \qquad (6)$$

Other properties of this entropy measure, such as those for dependent schemes, can be found, for example, in Khinchin paper [16]. For the purpose of our work, we will only consider mutually independent systems.

IV. Cellular Automaton (CA) And CA Entropy

The theory of cellular automata was initiated by John Von Neumann in his seminal work Theory of Self-Reproducing Automata [17]. It can produce complex phenomenon by simple cell and simple rules, which has the ability to model and simulate the complex system. Since the nineteen eighties, as the evolution of computer technology and the progress of science, cellular automaton theory gets in-depth researched and is widely applied in economic, transportation, physical, chemical, artificial life and other complex systems [9] [10] [11].

A cellular automaton consists of a regular grid of cells, each in one of a finite number of states, such as Black and White. The grid can be in any finite number of dimensions. For each cell, a set of cells called its neighbour (usually including the cell itself) is defined relative to the specified cell. An initial state (time t=0) is selected by assigning a state for each cell. A new generation is created according to some fixed rules that determine the new state of each cell in terms of the current state of the cell and the states of the cells in its neighbour.

In this work, we model the cloud scheduling system's behaviour as a cellular automaton (CA), specifically as a one-dimension CA network, and then calculate the CA entropy to measure the reliability degree of such complex system under different scheduling rules and resource allocation strategies. In this way, the collection of cells that composes the CA consists of a number of cloud resources that are rented for running the project (Each cell of CA corresponding to a cloud resource). The CA rules in our work are described as selected scheduling algorithms as follows:

- **First Come, First Served (FCFS)**: Tasks are executed according to the sequence of task submitting. The **first come** task will be scheduled on the available resource **first** as soon as it is submitted and then removed from the queue.
- **Min-Min**: All the tasks in a project will be ordered by their computing demands first. The task with the **minimum** computing demand will be scheduled first on the available resource which the completion time is **minimum** and then removed from the queue.
- **Max-Min**: All the tasks in a project will be ordered by their computing demands first. The task with the **maximum** computing demand will be scheduled first on the available resource which the completion time is **minimum** and then removed from the queue.

Each resource gets two performance states: Low Productivity (LP) and High Productivity (HP), which are correspondingly showed as Black and White in a CA grid map. The state of a resource is determined by its performance ratio under specify scheduling rules. The performance ratio of a resource (RPR) is calculated as follow:

$$RPR = \frac{The\ completion\ time\ for\ all\ its\ assigned\ task}{The\ completion\ time\ of\ the\ project(Makespan)} \qquad (7)$$

If the RPR of a resource is over 50%, then it is in High Productivity state, otherwise it is in Low Productivity state.

Reliability is one of the basic characteristics of complex system, which changes with system evolution. For cloud scheduling system, as one resource of it suffered enough power (Such power may cause by internal local activities or external force) strikes, it will fall into low productivity state or at the worst case it breaks down, this is called the resource collapse.

The collapse resource will influence the productivity state of all other resources and may cause them collapse as well, which lead the scheduling system progress out of order and into a disorder/chaos condition. Along with the increase in the number of collapse resources, hierarchical expansion, will eventually lead to the collapse of the whole scheduling system. Thus, the scheduling system is failed to deliver the project as original planned. We conclude that:

If a system is in order condition, is more reliable, or vice versa. The reliability can be measured by the disorder degree, thus Average Resource Entropy (ARE), of a system.

To evaluate the reliability of scheduling system in CA, we decrease the computing capacity of one resource by 1% for each time step, with a total of 100 time step, which simulates a resource from full computing capacity till break down. The whole scheduling system's evolution pattern is generated and represented by CA grids. Fig. 1 shows some examples of grid pattern generated by CA for running a project consists of 100 random tasks by FCFS algorithm with different number of allocated resources.

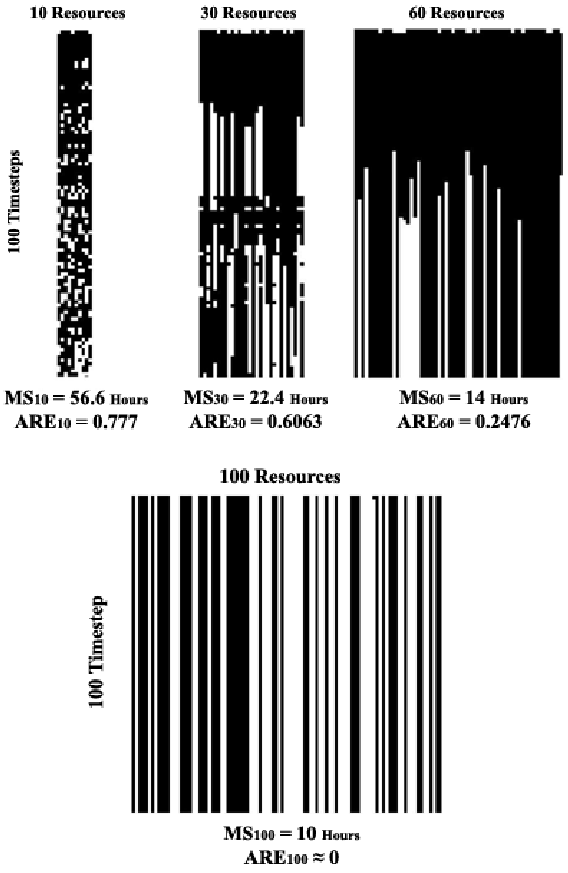


Fig. 1. Examples of Grid Pattern Generated by Cellular Automaton

The Average Resource Entropy in CA can be calculated by:

$$ARE = \sum_{1}^{n}(-p_{LP} \times \log p_{LP} - p_{HP} \times \log p_{HP})/n \qquad (8)$$

Where n refers to the number of resources that rented for running the project, p_{LP} and p_{HP} refers to probability of Low Productivity State and High Productivity State for a resource respectively.

V. Cost-Efficient And Reliable Resource Allocation (CERRA) Model

In this section, a Cost-Efficient and Reliable Resource Allocation (CERRA) model for scheduling project on the cloud is proposed based on the CA Entropy.

The proposed model can be used to achieve the optimal resource allocation strategy by considering both cost-efficiency and reliability for running project on the cloud within deadline and cost budget. The main components and control flow of CERRA model are shown in Fig. 2.

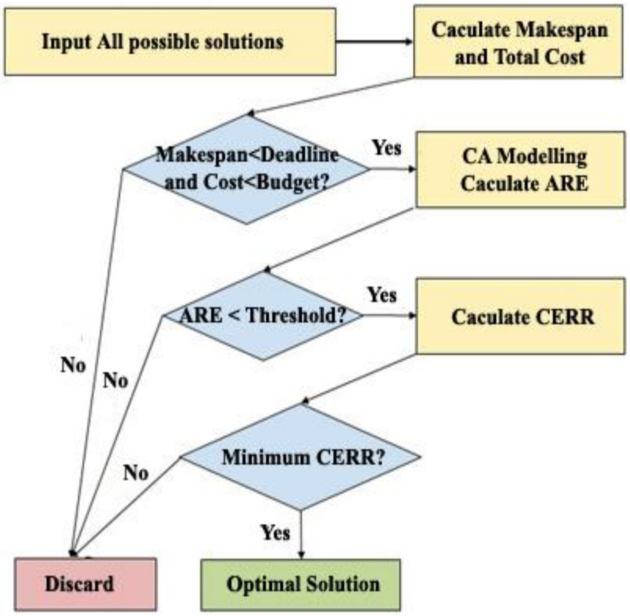

Input All possible solutions
Caculate Makespan and Total Cost
Makespan<Deadline and Cost<Budget?
Yes
CA Modelling Caculate ARE
ARE < Threshold?
Yes
Caculate CERR
No
No
No
Minimum CERR?
Yes
Discard
Optimal Solution

Fig. 2. Flow Diagram of CERRA Model

The optimal resource allocation solution selected by CERRA model meets the following condition:

- Meeting project deadline and within cost budget
- Under the reliability threshold that user prefer
- With the minimum Cost-Efficiency and Reliability Rate (CERR)

Where the Cost-Efficiency and Reliability Rate (CERR) is calculated by Formula:

$$CERR = \frac{MS_n \times \sum_{i=1}^{n} cp_i + ARE_n \times (MS_{n-1} - MS_n) \times \sum_{i=1}^{n} cp_i}{Cost\ Budget} \qquad (9)$$

Where n refers to the number of rented resources to run the project, MS refers to the project's completed time, cp refers to the cost price of a resource and ARE refers to the Average Resource Entropy.

VI. Experiment and Result

We implemented the proposed CERRA model under Matlab environment and simulated with three basic cloud scheduling algorithm, First Come First Served Algorithm (FCFS), Min-Min Algorithm and Max-Min Algorithm.

A. User Case 1- Simple Project Consists of 10 Random Tasks

A project consists of 10 tasks with random computing demand is listed in Table I. A maximum of 10 cloud resource units are available to be rent for running the project. The type of cloud resource units is M1 Small Instance which is based on Amazon EC2 instance types [17]. The specification of M1 Small Instance is shown in Table II. And the project requirements are shown in Table III.

TABLE I. PROJECT TASK SPECIFICATION

Task Specification	Task ID									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Computing Demand (Hours of 1 ECU)	12	3	7	9	15	24	10	1	2	4

TABLE II. CLOUD RESOURCE TYPE SPECIFICATION

Resource Type	Resource Specification		
	Computing Capacity	***Price***	***Available***
M1 Small Instance	1 ECU	$0.115 / Hour	10 Units

TABLE III. PROJECT REQUIREMENTS

Project Requirements		
Deadline	***Cost Budget***	***Average Resource Entropy (ARE) Threshold***
Makespan<35 Hours	Resource Cost<$20	ARE < 0.4

The experiment results for evaluating three selected scheduling strategies (FCFS, Min-Min, Max-Min) on all possible resource allocation are shown in Fig. 3, Fig. 4 and Fig. 5.

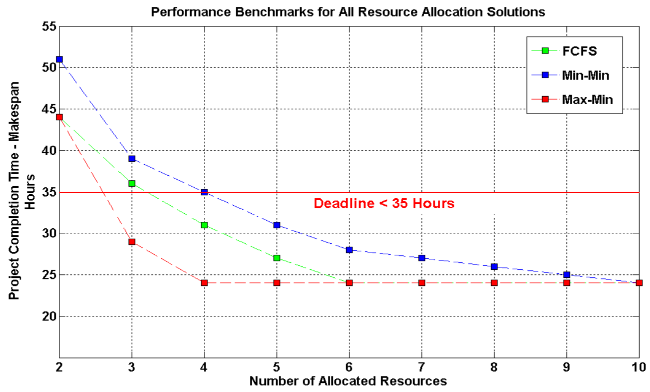


Fig. 3. Performance Benchmarks for All Resource Allocation Solution (10 Tasks)

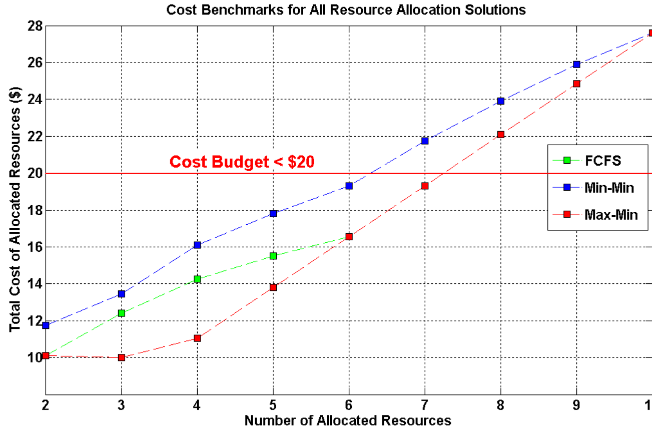


Fig. 4. Cost Benchmarks for All Resource Allocation Solution (10 Tasks)

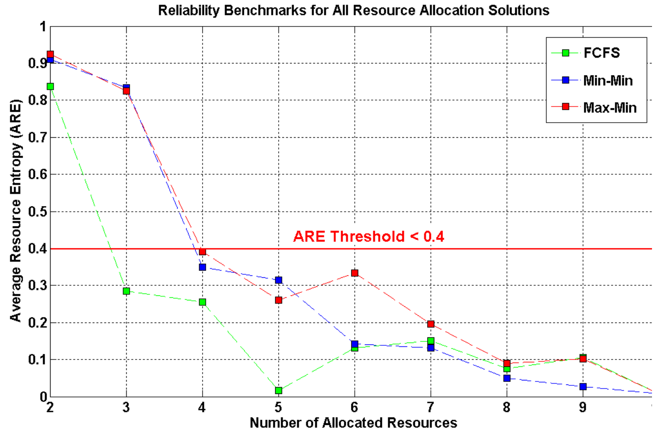


Fig. 5. ARE Benchmarks for All Resource Allocation Solution (10 Tasks)

- **Performance Benchmark:**

In general, the makespan of the three scheduling strategies for the project decrease as more resources are rented. However, over a number of resources, e.g. 4 resources for Max-Min and 6 resources for FCFS, any resources that newly invested do not improve the system's performance. With over renting of 5 resources, the improvement is limited for all the scheduling strategies. General speaking, Max-Min strategy performs better than FCFS and Min-Min for most of the solutions. Solutions with less than renting 4 resources are discarded because of failed to meeting the deadline, except for the solution that was allocated 3 resources with Max-Min scheduling strategy.

- **Cost Benchmark:**

In most case, cost of the project linearly increases as more resources are rented. Except for the solutions under Max-Min strategy, the cost for renting 2, 3and 4 resources is similar. Under the cost budget restriction, most of the solutions with renting more than 6 resources are discarded.

- **Reliability Benchmark:**

In general, adding a resource can improve the reliability of the system for all the three scheduling strategies. The reliability improvements for different scheduling strategies vary a lot. In the case of the number of renting resources equals the number of tasks, the project gets as many resources unit as it required and the Average Resource Entropy become zero for all the scheduling strategies.

In this case, scheduling system has zero entropy, indicating order and reliability. For this project, FCFS wins the reliability benchmark in most situations. Most of the solutions with renting less than 4 resources over ARE threshold are discarded.

At last, we calculate the Cost-Efficiency and Reliability Rate for all the resource allocation solutions; the CERR benchmark is shown in Fig. 6. We compare the CERR for all the remaining solutions that meet the project requirements as listed in Table III. With the Minimum CERR principle, the final result and detail performance of the optimal solution are shown in Table IV.

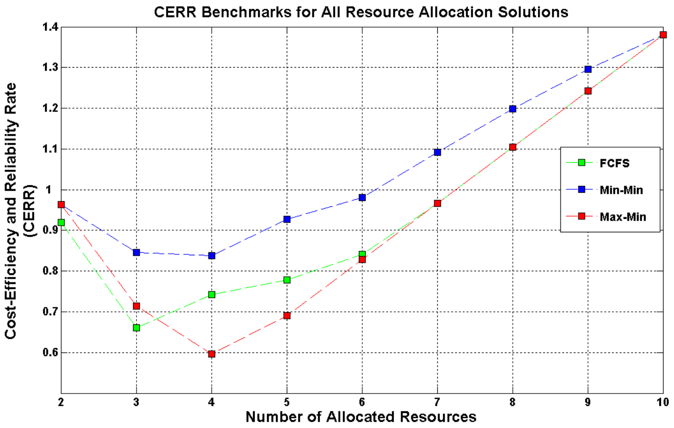


Fig. 6. CERR Benchmarks for All Resource Allocation Solution (10 Tasks)

As can be seen from the Table IV, solutions with allocating 4 resources for this project are optimal for three of the scheduling strategies. In most case, Max-Min scheduling strategy best fits the project with considering both cost-efficiency and reliability. However, decision maker that prefer to more reliable solution may choose FCFS scheduling strategy as it is more reliable than Max-Min.

TABLE IV. OPTIMIZE RESOURCE ALLOCATION SOLUTIONS (10 TASKS)

Solution Specification	**Solutions Ranking**		
	First Choice	***Second Choice***	***Third Choice***
Scheduling Strategy	Max-Min	FCFS	Min-Min
Rented Resources	4	4	4
Makespan	24 Hours	31 Hours	35 Hours
Cost	$11.4	$14.26	$16.1
Reliability (ARE)	0.3914	0.2550	0.3492
CERR	0.597	0.742	0.837

B. *User Case 2 – Complicate Project Consists of 100 Random Tasks*

In order to evaluate the robustness of the proposed CERRA model, a more complex project consists of 100 random tasks is presented as shown in Table V. The type of cloud resources and the project requirements are listed in Table VI and Table VII. As the project becomes more complicate, it becomes harder for a decision maker to seek out an optimal solution and make the project manageable. In this case, the reliability of scheduling system is an important factor that cannot be ignored which is related to the risk of failing to running the project as original planned. Thus, if the decision maker chooses a wrong scheduling strategy or resource allocation solution for a project, it will lead to dramatically increment of project cost or at the worse case failure of finishing the project within deadline. A suitable modelling and accurate measurement of the reliability of system is needed for planning such complicate and large project.

TABLE V. PROJECT TASK SPECIFICATION

Project Task Specification	
Total Number of Tasks	100
Total Computing Demand (Hours of 1 ECU)	5164
Maximum Computing Demand (Hours of 1 ECU)	100
Minimum Computing Demand (Hours of 1 ECU)	1
Probability distributions used for creating randomly generated tasks	Normal Distribution: Many middle size tasks, and fewer big and small tasks were contained in the project

TABLE VI. CLOUD RESOURCE TYPE SPECIFICATION

Resource Type	**Resource Specification**		
	Computing Capacity	***Price***	***Available***
M1 Small Instance	1 ECU	$0.115 / Hour	100 Units

TABLE VII. PROJECT REQUIRENMENTS

Project Requirements		
Deadline	***Cost Budget***	***Average Resource Entropy (ARE) Threshold***
Makespan<200 Hours	Resource Cost<$800	ARE < 0.4

As Fig. 8 shows, the performances of the three scheduling strategies are quite similar under different resource allocation solutions. However, the costs for different scheduling strategies vary a lot as shown on Fig. 9. Max-Min scheduling strategy wins the cost benchmark for most of the resource allocation solutions. It is no doubt that Max-Min is the selected optimal cost-efficient strategy for this project.

From Fig. 10 we can see the reliability of the system under Max-Min strategy acts like random walk as the number of allocated resources increase. At the point of 30 resources are allocated, the reliability of the system is greatly improved. After that point, the average resource entropy (ARE) of the system increases dramatically and reaches its highest peak at the point of 42 resources then fall back to more order state at the point of 45 resources. In overall, the ARE curve oscillate largely and irregularly until reach the point of 60 resources. In most traditional way, such reliability of a scheduling system is hard to be modelled and measured, which result in being ignored by the decision maker, especially for planning large and complicate projects. With our proposed CERRA model, the above problem can be solved by the quantitative measurement of average resource entropy in the system.

Fig. 7 shows the CERR benchmark for all the resource allocation solutions for the project. Table VIII list the comparisons of several near-optimal resource allocation solutions for running the project under the same Max-Min strategy.

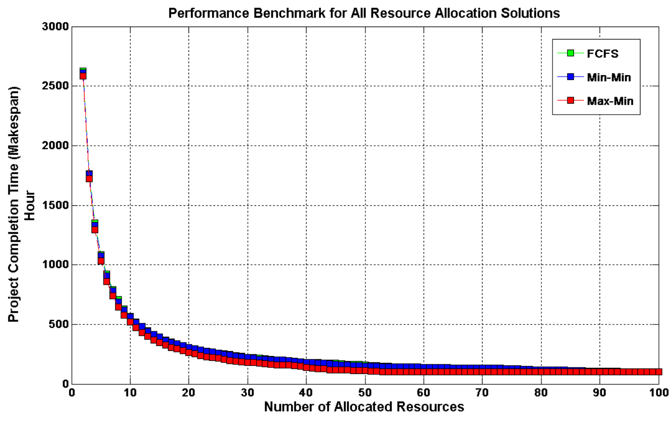


Fig. 8. Performance Benchmarks for All Resource Allocation Solutions (100 Tasks)

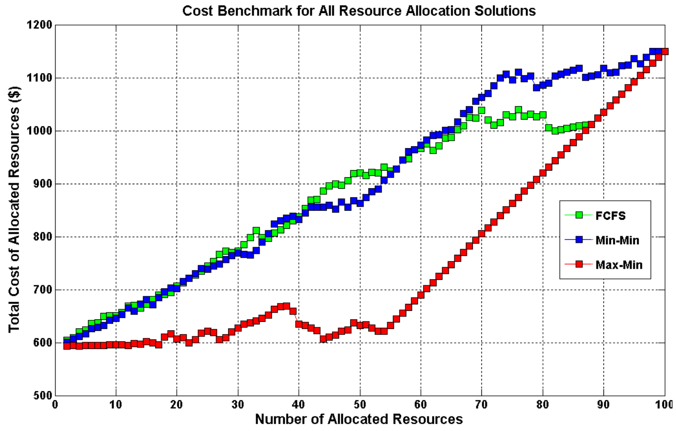


Fig. 9. Cost Benchmarks for All Resource Allocation Solutions (100 Tasks)

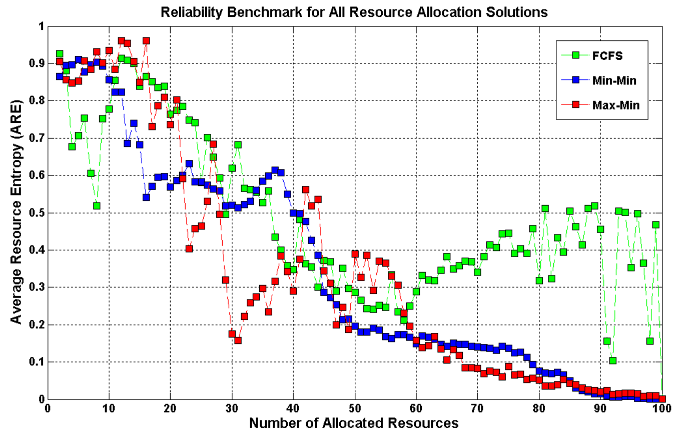


Fig. 10. ARE Benchmarks For All Resource Allocation Solutions (100 Tasks)

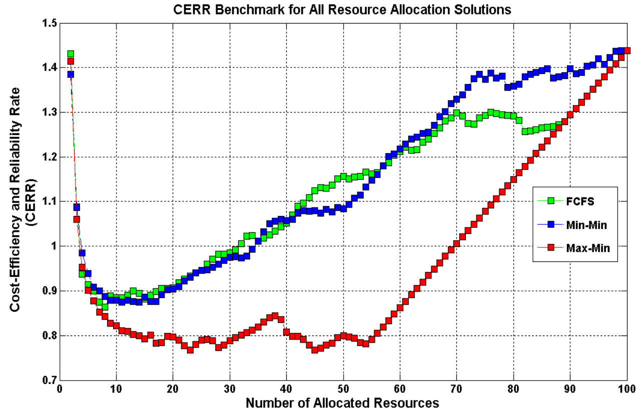


Fig. 7. CERR Benchmarks for All Resource Allocation Solution (100 Tasks)

TABLE VIII: COMPARISON OF DIFFERENT RESOURCE ALLOCATION SOLUTIONS UNDER MAX-MIN SCHEDULING STRATEGY

Max-Min Scheduling Strategy	**Resource Allocation Solutions Comparisons**				
	1	***2***	***3***	***4***	***5***
Rented Resources	23	30	38	44	45
Makespan (Hour)	229	182	153	120	118
Cost ($)	605.71	627.90	668.61	607.20	610.65
Reliability (ARE)	0.404	0.175	0.384	0.536	0.343
CERR	0.768	0.787	0.844	0.779	0.767

Form Table VIII, some observations were drawn.

- Observation 1: With minimum CERR principle, solution 5 is selected as the optimal solution for running the project.
- Observation 2: Although the solution 4 is discarded because its reliability degree (0.536) is over ARE threshold (ARE<0.4). It is still a near-optimal solution that performs close to solution 5.
- Observation 3: Compare solution 3 with solution 2 and 4, we can see solution with an allocation of 38 resources for the project result in disproportional return on investment.
- Observation 4: The CERR value of solution 1 is close to solution 5 with similar cost and reliability degree but huge performance difference. Since we measure the CERR by considering strictly "meeting the deadline" only, excluding the saving time cost of meeting the deadline. In the future, this factor should be considered in our CERRA model.

In summary, the proposed CERRA model is capable of providing useful information and quantitative measurement for aiding the decision maker to achieve a Cost-Efficient and Reliable solution for planning projects on the cloud.

VII. Conclusion

Resource Allocation in cloud scheduling system is a complex problem, the solution of which requires suitable modelling and complex optimization calculations. The CERRA model proposed in this paper puts forward an optimization method that is different from the traditional approach. It is one that is based on Cellular Automaton Entropy, based on minimizing the CERR of a scheduling system, which indicates both cost-efficiency and higher level of reliability resource allocation solution thus a more manageable project. The proposed model has been applied to aid decision maker for planning project on the cloud. The experiments help demonstrate how the CERRA model can be implemented and interpreted, and how the CA Entropy-based solution can be introduced in a project manager's decision-making process. The experiment result shows that the proposed model is able to achieve both cost-efficient and reliable resource allocation solution for running project on the cloud by solving the questions when planner making a decision:

- How many resources do I need?
- How should I schedule the project on the resources?
- Is it such solution cost-efficient and reliable?
- Giving a collection of solutions, which one is better for different requirements?

VIII. Future Work

Since the approach of applying Cellular Automaton Entropy to analysis the cloud scheduling system in this paper is the first attempt in the related literature. Many problems may arise, and many issues remain open. Future work should further examine and expand the entropy method presented in this paper: (1) we would like to determine if our model can be generalized for other types of complex scheduling strategies, e.g. Genetic Algorithm (GA), DAG scheduling algorithm, Simulated Annealing (SA), Ant colony optimization (ACO) Algorithms; (2) We would also like to study the optimization problem when allowing for the heterogeneous types of cloud resources and the dynamic adjustment of the number of cloud resources during run-time; (3) Also we are currently developing our model as a web service to help user in improving their resources renting strategies in the Cloud environment; (3) Finally, we plan to research on its application in other similar real-world problems, e.g. Staff Shifting Management, Calling Centre Scheduling, Traffic Routing, Manufactory Production.

References

[1] Christodoulou, Symeon, Georgios Ellinas, and Pooyan Aslani. "Entropy-based scheduling of resource-constrained construction projects." *Automation in Construction* 18.7 (2009): 919-928.

[2] Hermenier, Fabien, et al. "Entropy: a consolidation manager for clusters." Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments. ACM, 2009.

[3] Gan, H-S., and A. Wirth. "Comparing deterministic, robust and online scheduling using entropy." International journal of production research 43.10 (2005): 2113-2134.

[4] Botón-Fernández, María, Francisco Prieto Castrillo, and Miguel Vega-Rodríguez. "Nature-inspired algorithms applied to an efficient and self-adaptive resources selection model for grid applications." Theory and Practice of Natural Computing (2012): 84-96.

[5] Liu, Jiansheng, et al. "Research on Measurement Entropy-Based of Equipment Management Complexity and Its Application in Production Planning." Intelligent Robotics and Applications (2008): 604-611.

[6] Braun, Tracy D., et al. "A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems." Journal of Parallel and Distributed computing 61.6 (2001): 810-837.

[7] Gutierrez-Garcia, J. Octavio, and Kwang Mong Sim. "A family of heuristics for agent-based elastic Cloud bag-of-tasks concurrent scheduling." *Future Generation Computer Systems* (2012).

[8] Dong, Fangpeng, and Selim G. Akl. "Scheduling algorithms for grid computing: State of the art and open problems." School of Computing, Queen’s University, Kingston, Ontario (2006).

[9] Toffoli, Tommaso, and Norman Margolus. Cellular automata machines: a new environment for modeling. MIT press, 1987.

[10] Wolfram, Stephen. "Universality and complexity in cellular automata." Physica D: Nonlinear Phenomena 10.1 (1984): 1-35.

[11] Langton, Chris G. "Computation at the edge of chaos: Phase transitions and emergent computation." Physica D: Nonlinear Phenomena 42.1 (1990): 12-37.

[12] LEON, O. "Local activity is the origin of complexity." International journal of bifurcation and chaos 15.11 (2005): 3435-3456.

[13] Matthews, Robert AJ. "The science of Murphy's Law." SCIENTIFIC AMERICAN-AMERICAN EDITION- 276 (1997): 88-91.

[14] Lambert, F.L.: The Second Law of Thermodynamics. http://www.secondlaw.com, 2005.

[15] Amazon EC2, http://aws.amazon.com/ec2/, 2013.

[16] Khinchin, A. Ya. Mathematical foundations of information theory. Dover Publications, 1957.

[17] Von Neumann, John, and Arthur W. Burks. "Theory of self-reproducing automata." (1966)

Analysis of an Automatic Accessibility Evaluator to Validate a Virtual and Authenticated Environment

Elisa Maria Pivetta, Carla Flor, Daniela Satomi Saito, Vania Ribas Ulbricht
Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brazil

***Abstract*—This article's objective is to analyze an automatic validation software compatible with the guidelines of Web Content Accessibility Guidelines (WCAG) 2.0 in an authenticated environment. To the evaluation it was utilized as a test platform the authenticated environment of Moodle, which is an open source platform created for educational environments. Initially, a brief conceptualization about accessibility and the operation of these guidelines was described, and then the software to be tested was chosen: the WAVE. In the next step, the tool's operation was valued and the study's analysis was made, which allowed the comparison between the testable errors of WAVE with the guidelines of WCAG 2.0. As the results of the research, it was concluded that the tool WAVE obtained a good performance, even though it did not include several guidelines of WCAG 2.0 and did not classified the results within the accessibility's principles of Web Accessibility Initiative (WAI). Also showed itself more adequate to developers than to common users, which have no knowledge of Web programming language.**

Keywords*—automatic validation tool; WCAG 2.0; accessibility; WAVE)

I. INTRODUCTION

Web accessibility refers to the capacity of people, regardless of their skills, to perceive, understand and execute activities of navigation and interaction, as well as create contents in web [1]. The accessibility problems in the web affect people with deficiency, being it visual, motor, hearing, cognitive, language, neural system disturbs and others. Although not only people with deficiency need accessibility. Elderly, temporary deficiencies and people in general need accessible environments.

In Brazil, the decree 5.296, published in December 2004, makes mandatory the accessibility in websites of the public administration for the use of people with deficiencies with the objective of ensure full access to the contents available in the web[2].

The inclusion of people with deficiency in the educational, professional and social ways, besides of being mandatory by the Brazilian legislation, is also a social justice act, and provides an independency perspective to the individuals when communication and interaction barriers are diminished.

According to a research made in 2012 by W3C.br/NIC.br, only 2% of government web pages are accessible [3]. In international level, the World Wide Web Consortium [1] has an accessibility working group, which was created to discuss and plan acts in favor of the accessibility in web.

In scope of online education, the IMS Global Learning Consortium [4] conceptualizes accessibility as "the ability of adjust the learning environment to the necessities of all students". This accessibility can be determined by the flexibility of the environment and the availability of contents and alternative activities. In order to guarantee an effective access to Distance Education, the decree n. 5.622 of December 2005, has in the II item of the Article n. 13, that the pedagogical projects of distance courses and programs must offer appropriated treatment to students with special necessities [5]. An appropriated treatment also implies the offer of a technologic structure; in other words, an accessible Virtual Environment of Education.

The Virtual Environments of Education are systems based on a collaborative approach to the creation, application and management of courses that use the Internet. By having these characteristic, they present to have inclusive elements and they are largely utilized also as a support to the presential education. An example of a Virtual Environment of Education is the Moodle platform (Modular Object-Oriented Dynamic Learning Environment), which is the most known and utilized open-source environment in the world [6].

The Web Accessibility Initiative presents guidelines and recommendations to provide access and egalitarian opportunities to people considering the several types of skills in digital environments. Within these guidelines, WAI recommends as a preliminary revision the use of accessibility evaluation tools to identify possible problems that occur in a website. There are several ways to verify the accessibility in Virtual Environments of Education. One of them regards the automatic evaluators, which are softwares that test virtual environments by analyzing the code to verify if those are in conformity with the accessibility guidelines selected to the inspection. Nowadays, the prepositions of WAI, more specifically the Web Content Accessibility Guidelines 2.0 [7], are important references when discussions about web accessibility are raised [8].

Within the accessibility context, searching for an automatic tool that proposes evaluate authenticated environments, this article had as an objective to analyze one of these tools in relation to the virtual environment of education Moodle.

The selected software to the analysis is nominated of WAVE [9] and the choice method warned the observance of the guidelines and recommendations of WCAG 2.0, the philosophy of free code and the validation of an authenticated environment.

II. Web Accessibility Initiative (WAI)

The Web Accessibility Initiative has as proposal to present guidelines and recommendations to provide accessibility [8]. In order to support evaluators, developers and authors of contents in the production of accessible and usable contents by deficient people, WAI articulated the elaboration of the Web Content Accessibility Guidelines (WCAG), which nowadays can be found in the 2.0 version.

According to Reid e Snow-Weaver [10], the referred document has as one of its biggest objectives to describe the requirements to the accessibility of web contents in a neutral language of technologic and in a way that it can be applicable in any technology W3C or not, as CSS, SMIL, SVG, PDF or Flash, in addition to HTML and XHTML.

The accessibility guidelines were built based on four basic principles to a website [11]:

- Perceptible – the information and functionalities must be presented in a way that users can percept them
- Operable – the interactive functionalities must be available to users in a way that users can operate them
- Comprehensible – the information and functionalities must be clear to the understanding of users
- Robust – the contents must be robust enough to be reliably interpreted by a vast variety of agents, including assistive technologies.

Referring to the four principles, there is a list of twelve guidelines with orientations for the content to be accessible for the biggest amount of people. In the bottom of ever guideline there are success criterion that describe specifically what should be achieved, in order to fulfill the rule. All the success criteria of WCAG 2.0 are written as testable criteria to objectively determinate if the content satisfies those criteria. While some tests are automated by utilizing evaluation software programs, others need human testers in a part or in the whole test.

The guidelines are available in WCAG 2.0 [11], where the accessibility is identified in the following levels:

- Level “A” of conformity: is the minimum criterion of conformity, where all the success criteria categorized as A are satisfied
- Level “AA” of conformity: all the success criteria categorizes as A and AA are satisfied
- Level “AAA” of conformity: all the success criteria categorized as A, AA and AAA are satisfied

It is important to say that the success criteria adopted to the conformity levels are determined having as measure the difficulty level that they present to deficient people, when compared to other publics (by the committee’s point of view).

Besides the principles, WAI has non-testable recommendations, but those are ones that give framework global objectives to help the understanding of the success criteria and implement techniques in a better way.

III. Software of Automatic Evaluation

An automatic evaluation program, usually called validator, evaluator or online validator, is a set of tools that evaluate the content of a website according to a set of standards that determinate the accessibility level of the document. To do so, it needs to detect the code of a web page and analyze its content based on guidelines and accessibility recommendations such as W3C [1] and the Section 508 [12].

The validator helps to verify if the analyzed interface was developed by using the web standards of accessibility. In general, these programs are available on the internet by being commercialized or by free distribution, and several differences between them are pointed. Referring to the use of guidelines of W3C, several of them attend only the 1.0 version of WCAG. However, the current guidelines can be found in the document WCAG 2.0 [11].

The evaluations made thru validators are usually fast, but not capable to identify all the accessibility aspects. In general, the utilized tools make the verification based on the W3C recommendations, even if some of them are capable to analyze the submitted document deeper than others. Considering the different criteria that can be adopted to the validation of each one of the accessibility recommendations, the validators present some differences in relation to the answers, warnings and identified problems. According to Faulkner and Arch [13], the automated tools:

- Verify the code’s syntax;
- Identify real accessibility problems;
- Identify some potential problems;
- Identify pages that contain elements that might cause problems;
- Search for known standards.

IV. Why WAVE?

There are several accessibility validation softwares available on the internet. Initially it was selected the ones indicated by WAI, although by the moment of this research the WAI’s list was outdated, not having indications of any automatic tool that validates the conformity of a document with WCAG in the 2.0 version, but only with the WCAG 1.0.

Then it was considered only the softwares based on the WCAG 2.0 guidelines, considering that some were identified by Al-Khalifa et al [14] and others identified by the authors of this article. Another prerequisite to the choice was that the softwares attend to the open source philosophy, or at least that did not present a cost for acquisition. The table 1 presents the tools that were found and selected.

TABLE I. AUTOMATIC EVALUATION TOOLS

Software	Description	Levels of Conformity
AccessMonitor [15]	Developed by UMIC (Agency to the Society of Knowledge). It had as a starting point the accessibility evaluation tool eXaminator to WCAG 1.0. It emits an accessibility report and a synthesis of the results with an index, that is a valuation unit which the final result synthesizes and quantifies the level of accessibility achieved.	A, AA, AAA
AChecker (Public)[16]	Developed by the Adaptive Technology Feature Centre from the University of Toronto. It presents the results in three categories: known problems, probable problems and potential problems.	A, AA, AAA
ASES 2.0 [17]	Avaliator and Simulator of Accessibility of Sites- its objetive is to provide instruments that make the adoption of accessibility in government sites possible. According to ASES [17], it has tools that evaluate the conformity according to guidelines of WCAG 2.0 and e-MAG 3.0 [18].	A, AA, AAA
TAW3 [19]	It evaluates web pages and stand-alone Java applications. It presents the result in three categories: problems, advertences and non-verified. Based on the accessibility fundaments proposed by WAI.	A, AA (free version) AAA (commercial)
WAAT [20]	Web Accessibility Assessment Tool- Java application developed by the EU FP7 ACCESSIBLE project. Based on the accessibility fundaments proposed by WAI.	A, AA, AAA
WAVE [9]	Web Accessibility Evaluation Tool- avaliable by WebAIM.Provides four kinds of report: mistakes, features and warning; analysis of the page's structure, identifying the sequency of navigation; presentation of the page in the only text mode and, finally, identification of the headers of the page.	A, AA, AAA
Worldspace FireEyes [21]	Conceived as add-on of the Fifefox navigator, tests static and dynamic contents.	A, AA

Source: from the authors

The selected softwares were evaluated by three specialists: two graduated and with a master degree in Computer Sciences; and one graduated in Design with a master degree in Engineering and Knowledge Management. The three evaluators are PhD students and participants of the researching group in digital accessibility.

By the first step of the process, the selected tools were tabulated having as requisite to make an accessibility evaluation based on the document WCAG 2.0. Because accessibility validation includes generically several accessibility problems and the main-public in question is deficient people, it is considered that to the achievement of an evaluation result with depth, the three levels: A, AA and AAA, were considered relevant in an application designated to this. It is important to say that the success criteria and the conformity levels adopted by the WAI guidelines are determined based on the difficulty level that deficient people present when compared to other publics (by the committee's point of view) [1]. Therefore, of this group, the tools Worldspace FireEyes and TAW3 were disregarded for evaluating only in two levels of conformity WCAG, which are A and AA.

The next step consisted in execute and test the remaining sofwares in relation to the Moodle environment. Some softwares did not execute correctly when the evaluation sceneries required an user authentication using username and password, such as happens in the Moodle configurations.

From the referenced softwares in Table 1, only two were successful in authenticated sceneries: WAVE and ASES. Others, like TAW3, AChecker and WAAT only presented the possibility of evaluation of these environments thru the option of file upload or copy of source-code. In this context, Pivetta, Saito and Ulbricht [22] execute in their work an evaluation of Moodle by utilizing the quoted options. The authors noticed that this evaluation strategy, for being and offline approach, thus a static analysis, the tools could not evaluate completely the codes that were sent, whereas the evaluated pages were making reference to style pages (CSS files) and JavaScript extern files. Even with limitations, the tools could identify a part of the accessibility problems in the evaluated code. The positive factor about these tools is that they present their reports classified within the four principles of WCAG 2.0, presented in the table 2. However within the methodology utilized to execute this study, none of these tools was selected.

TABLE II. MOODLE ANALYSIS WITH NON-AUTHENTICABLE AUTOMATIC TOOLD

Software	Perceptible	Operable	Comprehensible	Robust
TAW	2	1	-	-
AChecker	10	-	-	-
WAAT	222	2	6	7

Source: Pivetta, Saito and Ulbricht [22]

Next, the softwares WAVE and ASES were tested in the Moodle environment. ASES presented execution problems while the tests and, for that reason, this article includes only the evaluation of WAVE, keeping the evaluation report of ASES for a posterior work.

From this choice, a search in the CAPES[1] website was made, and also in the searching website Google.com, to verify the art state in relation to automatic tools evaluations. Some related works were found in the Google[2] website, such as an article by Faulkner and Arch [13], which refers to the evaluation of four tools of automatic validation. In this article, the WAVE tool is quoted, but not evaluated, being the commercial softwares the tests' main objectives of evaluation. For being an older article, the evaluation arguments regarded the guidelines of WCAG 1.0, document that was already overcome by the 2.0 version. Still, Faulkner and Arch [13] related other works of evaluation of automatic tools, but that

[1] http://www.periodicos.capes.gov.br.ez47.periodicos.capes.gov.br/ - access in 12/2012

[2] http://www.google.com.br – access in 02/2013

also refers only to WCAG 1.0, which differs of this proposal. Alexander and Rippon [22], however, utilized WAVE in their work to evaluate an academic website, not including properly an evaluation of the WAVE software, but its application to web environments evaluation.

V. Evaluation Of WAVE

WAVE is a set of accessibility evaluation web tools based on the guidelines of WCAG 2.0 [7] and Section 508 [12], of free access, developed by WebAIM [9]. The validator, instead of providing a technical report like most of automatic evaluation programs, shows the evaluation result on the web page that originated the tests by utilizing embedded icons and indicators that reveal the pages' accessibility. It allows evaluate an URL (Uniform Feature Locator) of a website, although in case the files are not publicly available on the internet, there is still the possibility of making the upload of the files to evaluation WAVE in the tool's website.

Another possibility presented is to copy the HTML code of the website chosen to execute the test and paste in the formulary available on the website. Besides, WAVE offers the download option of a toolbar in Mozilla Firefox, which is installed as a complement to the navigator.

Considering the deficiencies already discussed of an evaluation by upload of source-code, to test the Moodle environment internally the download of the toolbar WAVE Firefox was necessary. Already installed, the Moodle environment was executed with user and authentication password.

The WAVE Firefox toolbar allows evaluating web pages directly in the navigator and test environments that are protected by passwords. According WAVE [9], the toolbar evaluates the contents exhibited locally and dynamically, made from scripts or AJAX. It is composed by four tools that execute the verification of:

- Errors, features and alerts;
- Order of the structures;
- Only texts;
- Visualization of headers.

The toolbar contains other options, such as the option to disable styles, a link to the page that contains explanations about the accessibility icons and the option to "clean" evaluations that were already executed.

Every time a page is submitted to evaluation, colorful icons with different shapes appear as a result of what was evaluated. These are the identified categories:

- Red icons – indicate accessibility errors, which is, contain accessibility problems.
- Yellow icons – indicate alerts and, in this case, can be or not accessibility issues, but generally indicate an area where accessibility is, for several times, a problem, or that can be improved
- Green icons – indicate areas that contain elements with accessibility features and that the author must verify the accuracy.
- Light blue icons – indicate structural, semantic or navigation elements that can help in accessibility. These icons must also be verified.
- Trapezium icons – related to the images available on the website.

All the existing icons can be visualized in Fig 1. Also in the WAVE toolbar there is a link with explanations about the meaning and the recommended actions for each element.

To execute a WAVE report, it is necessary to select one of the four tools that compose the toolbar. The first test was with the option ERRORS, FEATURES AND ALERTS. The names of the items of the WAVE tool will always be referenced in capital letters to be differentiated from the text. In this first evaluation, the validator was applied to Moodle's homepage, where the authenticated user, in this case a student, can visualize courses, disciplines, his profile and other things. The screen referred to this first evaluation can be visualized in the Fig. 2, that shows how WAVE presents the results when submitted to validation of a website.

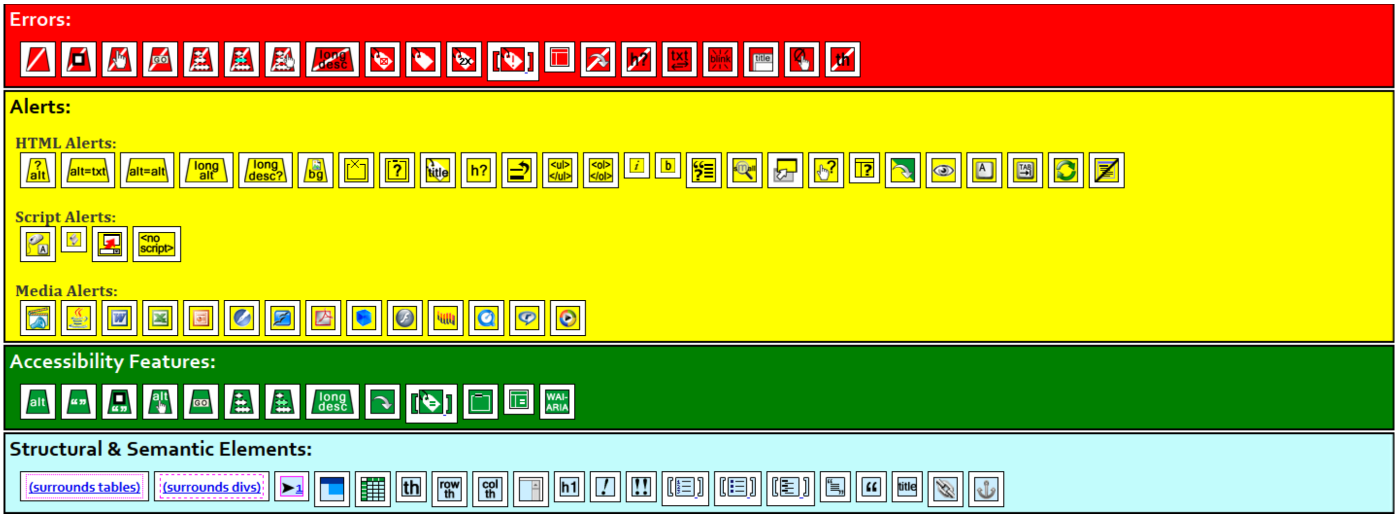
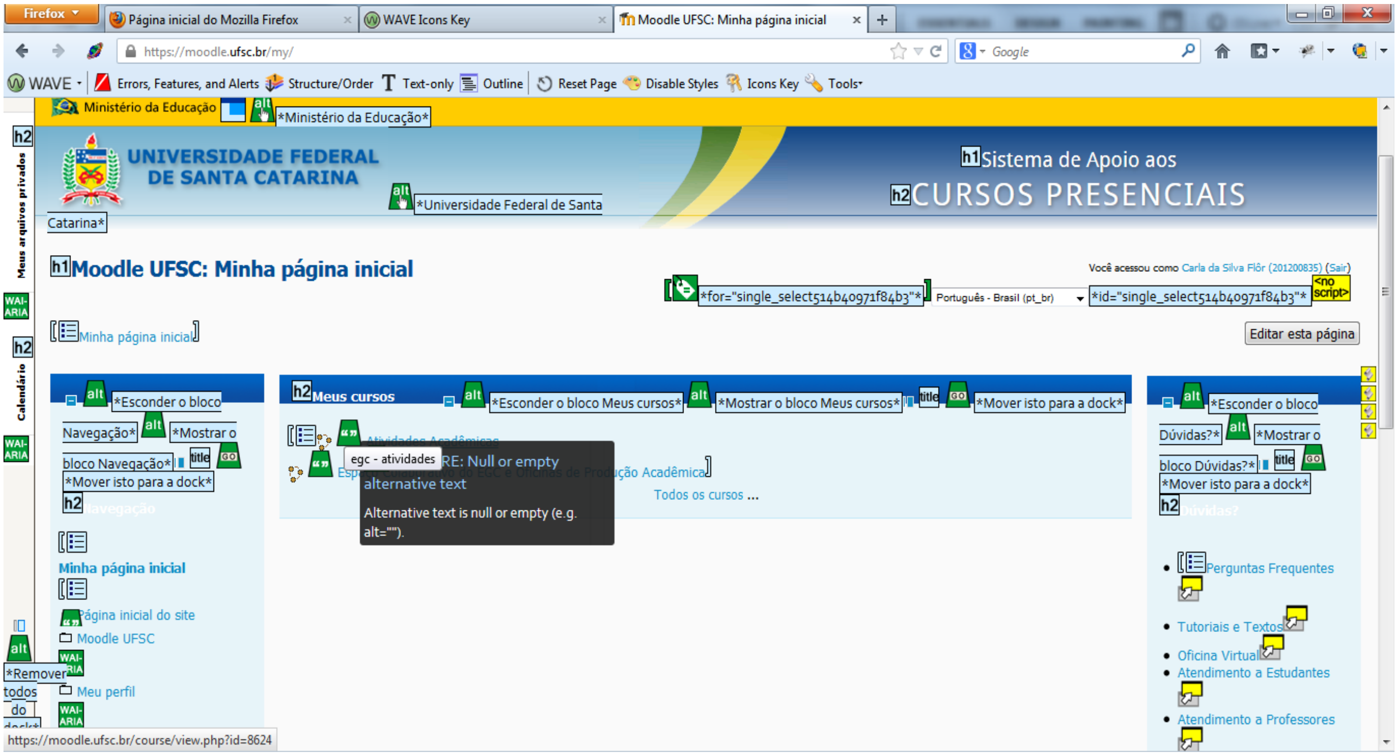


Fig. 1. Icons to accessibility indication (source: http://wave.webaim.org/icons)



RE: Null or empty alternative text

Alternative text is null or empty (e.g. alt="").

Fig. 2. WAVE test – ERROS, FEATURES, and ALERTS

The page is modified by the presence of accessibility icons in different areas of the website. Every time that the mouse cursor is positioned over one of these icons a brief description about the icon is presented, as you can see in a black rectangle in the Fig. 2. Or still, the user can see a detailed description of each icon in the option “Icons Key” that can be found in the right superior area of the page. The result of the tool in the homepage was showed thru the graphic presentation of icons:

- Yellow icons – alert about alternative texts to near images, links to new windows, CSS to occult content that were not read by screen readers, alternative texts for non-executed scripts on the navigator, JavaScript and others.
- Green icons – show existing accessibility features, like alternative contents for images and buttons.
- Blue icons – indicates non-enumerated lists and titles to the sections.

The evaluated page did not present any red icon, which is the icon that points errors, even when the use of CSS (Cascade Style Sheet) is disabled as WAVE [9] suggests.

As WAVE tests one page at a time, from the several screens tested, the error indication occurred only in pages that contained formularies. The registered errors were “image without textual alternative content” and “orphan labels”, in other words, without an associated entry. In the other pages the result was similar to the one obtained with the homepage.

According to WAVE [9], the functionalism of the WAVE 1.1.8 toolbar is also available in the tool menu, with the acceleration keys (ALT + T). It allows the accessibility of the keyboard to all the tool’s functions, even when the toolbar is not visible. In tests executed in three computers, those acceleration keys did not work. Usually the presentation of an underlined letter in the menu indicates its use as an acceleration feature, and when the underlined letter is pressed at the same time as the key ALT, its functionality is activated. In this case, ALT functioned normally to all the other options in the Firefox menu.

The next tool of WAVE is the STRUCTURE ORDER, which allows the visualization of the structural organization of the website. In this tool, the indicators show a reading sequence that corresponds to the order of navigation in the page. To determine if the reading and the order of navigation of the page make sense and are logical, the numbers must be followed. The tool also indicates the presence of lists, headers, tables and alerts in relation to the structure and functionality of these elements. The submission’s result can be observed in the picture 3.

In sequence, the presented tool is the TEXT ONLY. This tool provides the option to visualize only textual information of the page. The tool removes the page’s visual style and provides a verification of what is read by a screen reader, including alternative texts to images and bottoms. Besides, with this tool, other occult information to the user becomes visible, such as the “skip navigation” and “skip main menu” links.

The DISABLE STYLES tool has a similar effect. However, it is different from TEXT ONLY just for removing the page’s styles, maintaining the images and not showing the alternative contents to images and bottoms like TEXT ONLY.

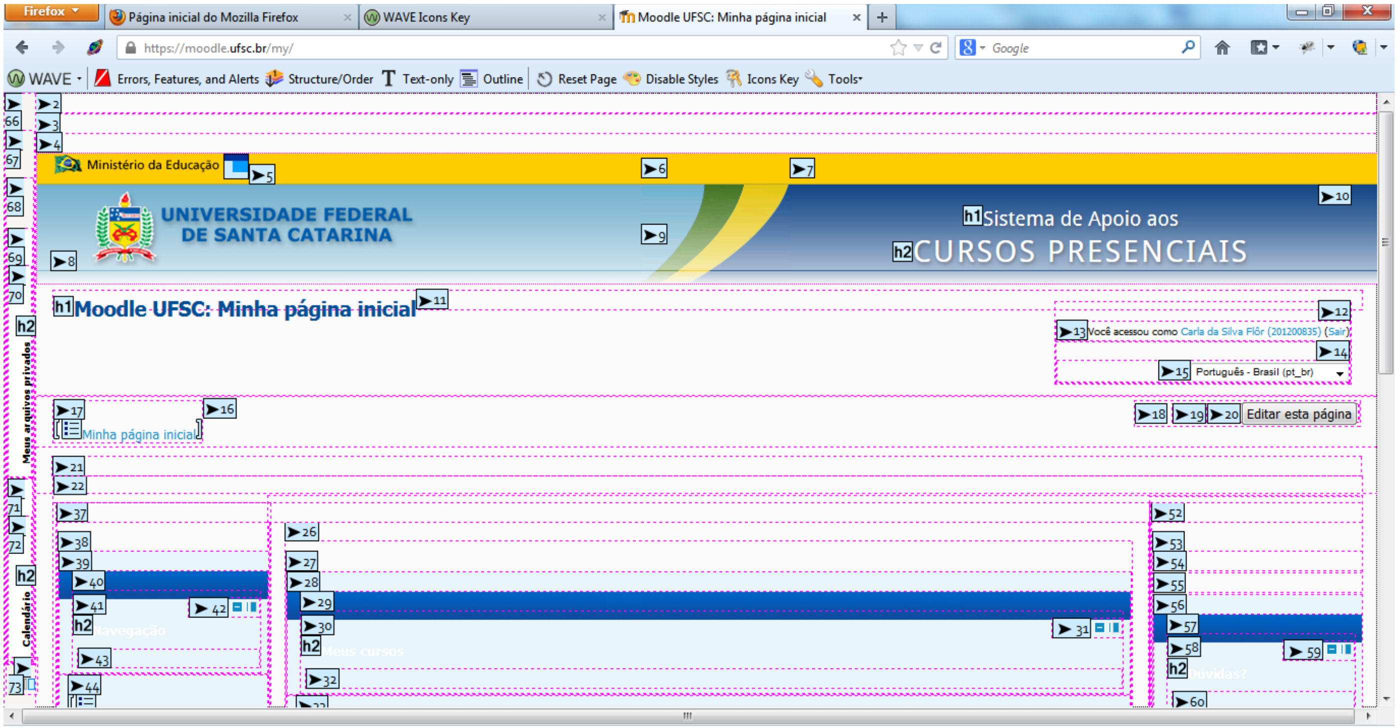


Fig. 3. Test with STRUCTURE ORDER

Lastly, the OUTLINE tool allows the visualization of headers and its levels, verifying if the structure is logical and adequate.

VI. Analysis of the Study

The WAVE, object of study of this work, was efficient when it comes to evaluation. However, it was observed that the results presented a few accessibility errors when compared with other analysis with the Moodle environment, such as the WAAT tool, described in Pivetta, Saito and Ulbricht [21]. Due to these results, it was made an analysis of tests and criteria utilized by WAVE and its tools to confront them with what is proposed by WCAG 2.0.

The W3C, thru WCAG 2.0, establishes as a base to the validation tests 60 testable success criteria that are included in the conformity levels A, AA and AAA. A great part of these criteria can be automatically tested to the presentation of results in reports to accessibility analysis.

In the WAVE analysis, it was verified the existence of 20 tests to accessibility evaluation in web pages in contrast with the 60 Success criteria of WCAG 2.0. From the 20 tests, 8 make reference to errors with image insertions, 4 make reference to errors of labels in formularies and only 8 test other accessibility errors of pages. These 8 are divided in:

- 3 errors to title: no title in tables, in page, in header;
- 2 errors to links;
- 1 error for having an HTML <*marquee*> (moving texts);
- 1 error to header of table without text.

Because these 20 errors are not identify within the four principles of WCAG 2.0 this work tried to analyze these errors and classify them within the principles. The table 3 and 4 shows the result of the classification.

TABLE III. WAVE ERRORS AND COMPARATIVE WITH THE SUCCESS CRITERIA WCAG 2.0

ERROR Wave	**Success Criterion WCAG 2.0**
ERROR: Missing alternative text;	1.1.1 (Non-text Content)
ERROR: Spacer image missing alternative text;	1.1.1 (Non-text Content)
ERROR: Linked image missing alternative text;	1.1.1 (Non-text Content)
ERROR: Image button missing alternative text	1.1.1 (Non-text Content)
ERROR: Image map missing alternative text	1.1.1 (Non-text Content)
ERROR: Image map area missing alternative text	1.1.1 (Non-text Content) 2.4.4 (Link Purpose (In Context)) 2.4.9 (Link Purpose (Link Only))
ERROR: Server-side image map	No match in WCAG 2.0. Reference to Section 508
ERROR: Invalid longdesc	1.1.1 (Non-text Content)
ERROR: Form label missing	1.1.1 (Non-text Content) 1.3.1 (Info and Relationships) 3.3.2 (Labels or Instructions) 4.1.2 (Name, Role, Value)
ERROR: Empty form label	1.3.1 (Info and Relationships) 3.3.2 (Labels or Instructions)
ERROR: Multiple form labels	4.1.1 (Parsing)

ERROR: Orphaned form label	1.1.1 (Non-text Content) 1.3.1 (Info and Relationships) 3.3.2 (Labels or Instructions) 4.1.2 (Name, Role, Value)
ERROR: Frame missing title	2.4.1 (Bypass Blocks) 4.1.2 (Name, Role, Value)
ERROR: Broken skip navigation link	2.4.1 (Bypass Blocks)
ERROR: Empty heading	2.4.6 (Headings and Labels)
ERROR: Marquee	2.2.2 (Pause, Stop, Hide: (Moving, blinking, scrolling))
ERROR: Blinking content	2.2.2 (Pause, Stop, Hide)
ERROR: <title> is missing or not informative	2.4.2 (Page Titled)
ERROR: Empty link	2.4.4 (Link Purpose (In Context)) 2.4.9 (Link Purpose (Link Only))
ERROR: Empty table header	1.3.1 (Info and Relationships)

Source: the authors

TABLE IV. WAVE CLASSIFICATION AND COMPARATIVE WITH WCAG 2.0

Principles	**WCAG 2.0**		**WAVE**
	Guidelines	***Success criteria***	***Errors that correspond to the success criteria WCAG 2.0***
Perceptible	4	22	2 (1.1.1 - 1.3.1)
Operable	4	19	6 (2.4.4 - 2.4.9 - 2.4.1 - 2.4.6 - 2.2.2 - 2.4.2)
Comprehensible	3	17	1 (3.3.2)
Robust	1	2	2 (4.1.2 - 4.1.1)
Total	12	60	11

Source: the authors

Analyzing the Tables 3 and 4 it is noticeable that eleven from all the errors that were treated by Wave are in the perceptible principle, but are equal to only two from the twenty-two success criteria to this principle (it occurs because each success criterion can be tested by more than one Wave error).

From the eleven errors in the perceptible level, nine are related to the 1.1.1 criterion (Non-text Content), while three (one error is in both criteria) are related to the 1.3.1 criterion (Info and Relationships). In the operable level, eight Wave errors correspond to six success criteria, and WCAG 2.0 foresee nineteen success criteria to this principle. By the same way, in the comprehensible level, three Wave errors are related to only one success criterion of this level, the 3.3.2 (Labels or Instructions), while WCAG 2.0 foresee seventeen criteria. The only totally satisfied principle is the lustiness, that has only two success criteria and both are testable by Wave errors.

Therefore, it is noticeable that Wave does not contemplate big part of the success criteria of Wave 2.0, covering only eleven from the sixty success criteria foreseen by WCAG 2.0.Besides, the WAVE tool does not categorize the errors according to the four principles of WCAG 2.0 for presenting a different report, composed by graphic icons within the page.

The report formats vary a lot, depending on the target-public, on familiarity with Web Design and Web accessibility standards. In case of web designers, developers and evaluators that know which better format answer its necessities, are able to choose one appropriated tool. According to WebAIM [24] the evaluation tools include six report formats:

- Based in text – errors listed by line number.
- Based in text – errors listed by linked line number, which is, links errors with the source-code.
- Based in text – errors listed within source-code.
- Based in text – errors listed within source-code and GUI (graphic user interface) – uses tables to show the users three pages in once: the report based on text, the graphic interface of the user of the webpage, and the error cases detached in the code.
- Graphic – based on icons.
- EARL – the EARL reports are a W3C attempt to standardize accessibility reports and help the users to compare the efficiency of the accessibility tools [1].

However, thru this graphic approach of icons, the developer can verify: alerts, accessibility features and where they can be found. To each identified item, WAVE has recommendations, such as:

- HTML alerts – 25 recommendations;
- Script alerts – 18 recommendations;
- Media alerts – 14 recommendations;
- Accessibility features – 13 recommendations;
- Structural and semantic elements – 20 recommendations.

The alerts are necessarily accessibility errors, although present HTML, scripts or media points that deserve more attention and that could be improved. As an example, a text alternative named “image” could be created; even though it is present in the code, it is not representative to the user because does not describe the image content.

On the other hand, the accessibility features highlight the present features for the developer to be able to verify if these are correct. The structure and semantic elements indicate the structure, the navigation and the semantic of the page, in a way that a correct read of the order and the hierarchy of the information can be made.

VII. CONCLUSION

This work evaluated automatic software of accessibility evaluation in relation to an authenticated virtual environment:

the Moodle. Softwares of this category present different approaches, shapes, characteristics and benefits, considering that some of them can present a large quantity of tests while others sub estimate the existing problems in a website. The choice of the ideal tool depends on a set of abilities and how the evaluator defines the responsibilities of the site that he wants to test. An important efficiency measure of an automated product is the capacity of produce results without the necessity of a more human interpretation. In this sense, were adopted the following validation criteria of the program about to be evaluated: freeware tools, the WCAG 2.0 guidelines, the possibility of evaluation of authenticated environments and the absence of execution errors in tests. Within the identified and selected programs, the one with better characteristics was WAVE.

WAVE's proposal indicates that the tool is appropriated to help web developers to make the available content more accessible. The software in question does not describe of a content is accessible or not, but helps the evaluator to verify accessibility aspects of this content. According to information from the site [9] the use of the tool demands experience of the evaluator user, which is, it is important that the person that is analyzing the site has the knowledge in computer sciences to a better understanding of the alerts and errors.

Considering that WAVE is a software that proposes the accessibility evaluation of websites, the tools appear to be a little limited. One example refers to the tools TEXT ONLY and OUTLINE, which are not very useful in a accessibility validation level, only help the developer to verify the structure of the site in the text more and in the header structure mode, respectively.

As an accessibility evaluation tool, it could be rearranged in a way to be in conformity with WAI guidelines, considering the four basic principles: perceptible, operable, comprehensible and robust, in a way to available error reports, alerts and other items classifying them within these principles. On the other hand, the used methodology, which is the icons insertion within the website is interesting due to the facility to identify the area and the accessibility item that is present or not.

Lastly, to provide accessibility tools is a great step to web accessibility. However, it is important to say that the use of tools to verify accessibility is only the first step. Besides, the evaluator must be warned about the tool's limitations and have knowledge of accessibility subjects and its implications to deficient people, in order to interpret the reports about signalization of alerts and errors.

REFERENCES

[1] W3C, Evaluation and Report Language (EARL) 1.0 Schema, 2011, http://www.w3.org/TR/EARL10-Schema/

[2] BRASIL, Decreto nº 5.296. Presidência da República - Casa Civil - Subchefia para Assuntos Jurídicos, 2004. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/decreto/d5296.htm.

[3] W3CGT, GT Acessibilidade. http://www.w3c.br/GT/GrupoAcessibilidade#w3c_inicio_conteudo

[4] IMS, Global Learning Consortium. IMS Access For All v2.0 Final Specification. http://www.imsglobal.org/accessibility/

[5] BRASIL, Decreto nº 5.622. Presidência da República - Casa Civil - Subchefia para Assuntos Jurídicos. 2005. http://portal.mec.gov.br/seed/arquivos/pdf/dec_5622.pdf

[6] Moodle, Modular Object-Oriented Dynamic Learning Environment. http://www.moodle.org

[7] WCAG 2.0, Web Content Accessibility Guidelines 2.0. http://www.w3.org/TR/WCAG20/.

[8] WAI, Web Accessibility Initiative. http://www.w3.org/WAI/

[9] WAVE, Web Accessibility Evaluation Tool. http://wave.webaim.org.

[10] L. G. Reid, A. Snow-Weaver, "WCAG 2.0: Web Accessibility Standard for the Evolving Web". Proceedings of the 2008 International Cross-Disciplinary Conference on Web Accessibility, W4A. Beijing, China: ACM Press, 2008.

[11] WCAG 2.0 – Web Content Accessibility Guidelines. http://www.w3.org/TR/WCAG/

[12] Section 508. https://www.section508.gov

[13] S. Faulkner, A. Arch, Accessibility Testing Software Compared, 2003. AusWeb03, University website accessibility revisited. http://ausweb.scu.edu.au/aw03/papers/arch/paper.html

[14] H. S. Al-Khalifa et al. "A Pilot Study for Evaluating Arabic Websites Usign Automated WCAG 2.0 Evaluation Tools", Proceedings of the 2011 International Conference on Innovations in Information Technology. Riyadh, Saudi Arabia: IEEE Computer Society, 2011.

[15] UMIC, http://www.acessibilidade.gov.pt/accessmonitor/

[16] ACHECKER, Web Accessibility Checker. http://achecker.ca/checker/index.php.

[17] ASES, Avaliador e Simulador para a Acessibilidade de Sítios. http://www.governoeletronico.gov.br/acoes-e-projetos/e-MAG/ases-avaliador-e-simulador-de-acessibilidade-sitios.

[18] EMAG30, e-MAG – Modelo de Acessibilidade de Governo Eletrônico – versão 3.0, 2011. http://www.governoeletronico.gov.br/biblioteca/arquivos/e-mag-3.0/download

[19] TAW, Accesibilidad de Sitios Web. http://www.tawdis.net/

[20] WAAT, Web Accessibility Assessment Tool. http://www.accessible-eu.org/.

[21] WORLDSPACE, Worldspace fireEyes. http://www.deque.com/products/worldspace-fireeyes.

[22] E. Pivetta. D. S. Saito and V. R. Ulbricht. "WCAG e a Acessibilidade de Ambientes Virtuais de Ensino Aprendizagem na Perspectiva de Alunos Surdos". Proceedings of II Conferência Internacional de Integração do Design, Engenharia e Gestão para a Inovação, 2012, IDEMi 2012. Florianópolis, 2012.

[23] D. Alexander, S. Rippon, "University website revisited", AUSWEB07, 2007. http://ausweb.scu.edu.au/aw07/papers/refereed/alexander/paper.html.

[24] WebAIM, "Accessibility Evaluation Tools" http://webaim.org/articles/tools/

Narrowing Down Learning Research:Technical Documentation in Information Systems Research

Convergence of research areas for customer learning of technical product functionalities

Thomas Puchleitner
Institute of Information Science and Information Systems
University of Graz / evolaris next level GmbH
Graz, Austria

Abstract—**learning how to use technical products is of high interest for customers as well as businesses. Besides product usability, technical documentation in various forms plays a major role for the acceptance of innovative products. Software applications partly integrate personalized learning strategies but late developments in information and communication technology extend these potentials to the non-software sector too. Mobile devices as smartphones allow the linking between physical and virtual world and are thereby eligible instruments for product learning and the application of adequate learning theories. Very few scientific publications accurately addressing the learning of product features and functionalities can be depicted. By applying a research profiling approach as a stepwise analysis of available publications, relevant learning paradigms and their corresponding scientific areas are depicted. As this research topic relates to marketing as well as information systems research the applied approach may also show beneficial for other interdisciplinary intentions.**

Keywords—problem-based learning; self-regulated learning; self-directed learning; product learning; customer learning; consumer learning

I. INTRODUCTION

It is said that the average European owns about 10.000 items [1]. These items are acquired because of personal needs and habits and later become part of daily routines and usages. Businesses compete to gain potential customer attention by applying various advertising strategies and techniques from simple printed and multimedia commercials to different kinds of less consciously strategies like guerilla marketing. Most of these strategies are based on the idea that the potential customer uses communication channels to fulfill communicational needs and therefore the same channel could be used for brand or product placements. Channels where marketing activities are placed have usually no relation to the product itself nor can the advertising message be timed with a potential usage scenario of the advertised product. Marketing activities while a customer is actually using a product, especially within the adoption process, allow the purposeful transportation of marketing messages.

Besides the product itself as carrier of such information, additional product information, manuals or instruction sets act as connector between company and customer. Customers see these as part of the product and therefore include them into product evaluations and current or future buying decisions [2]. Software products allow the integration of learning techniques

The mutual relation between the product itself and its corresponding additional documents shows especially high importance for technical products and especially for software applications, as these require deeper product knowledge. Studies show that various experts for Internet, communication and media even doubt the disappearance of additional documentation for electronic devices [3]. Usability research focuses on different aspects of product design [4] where technical communication in form of external documentations [5] fulfills the purpose of guiding the user through a product related situation. Redish [6] gives a personal and detailed insight into the relation. While printed versions of product documentation only allow information flow from product producer to customer, recent developments in communication technology and diffusion of required hardware provide more versatile ways of information flow for physical consumer goods too. Technologies like machine-to-machine communications, cyber-physical systems or technologies to bridge virtual and physical worlds as NFC provide manifold ways of information exchange.

Our ongoing research focuses on the transition from classic product documentation and information to the formation of a bidirectional one-to-one communication channel between businesses and customers. This is applied by personalized learning methods with high customer acceptance in the field of learning to use product functionalities. While customers profit due to faster and more convenient learning progresses, businesses gain insights not only into their products but also into customer requirements. This opens opportunities to analyze real usage scenarios and determine what functionalities are utilized and how customers apply them. Additionally, issues and mistakes users are recurrently facing while usage are spotted.

Thereby businesses are given opportunity to adjust products and services for improved customer experience and higher user satisfaction [7]. Businesses applying such methods are able to gain competitive advantages and open new potentials for various marketing purposes. This area of tension between product marketing and digitally enabled product learning demands an alternative angle of view on learning theories with high relevance for information systems research. E-learning is a contiguous discipline and field of research but focuses primarily on teaching aspects between teacher and student and thus does not cover aspects of marketing or product adaption.

A. *Technical communication in marketing and information systems research*

Technical communication includes both internal and external information regarding the product where technical documentation refers to documents and information that is handed specifically to the user [2]. External documentation therefore acts as an instrument of marketing by allowing customers to enhance their product experience due to the application of feasible learning approaches. Literature for learning and learning paradigms in technical documentation rarely covers impacts for customer satisfaction or buying behavior while marketing and information systems aspects chiefly focus on the design and usability of products and their corresponding documentations. Technology acceptance research as well as usability research define the factor of *easy-to-use* as crucial for positive product adoption [8]. Nielsen [4] and Davis [9] both describe the easiness to learn how to use a product has especially high impact on “easy-to-use”. Products that are easy to learn therefore lead to a competitive advantage, which shows relevance whenever a potential customer enters the product adoption phase. Studies confirm that usage lifecycle for consumer goods decreases for various product groups like cars [10] or mobile phones [11]. Product groups with shorter usage lifecycle are rebought in shorter intervals and undergo these adoption phase more often, which also implies additional learning effort for customers and therefore higher product switching costs.

A bidirectional information flow in learning frees potentials for both businesses and customers. While businesses gain additional insights and knowledge according their product usage and handling customers benefit from the thereby improved product support and faster learning progress. In other words: an application of appropriate learning techniques and systems when using products fosters user experience in adoption as well as usage phases. The change in communication behavior and new technical possibilities allow new ways of knowledge and information transfer, which are addressed in research areas related to e-learning or m-learning. While these topics focus on the aspects of teaching little research is known in the context of learning how to use a product. Also a clear definition of product learning in terms of keywords or research areas is missing and corresponding literature is widely spread. This interdisciplinary research includes topics from the fields of learning, marketing as well as information systems research with an absence of connecting links between them. The here applied approach bridges this gap by providing an opportunity to determine accurate scientific literature due to the identification of the most relevant connecting terms used in publications.

By applying such an approach a broader perspective is ensured and relevant publications in these areas are highlighted. Learning theory literature acts as a foundation as the awareness in human learning theories ensures the application of customer accepted product support.

B. *Research methodology*

Learning in a scientific context is wide spread and reaches from various learning techniques to modern e-learning topics. In chapter 2 we focus on relevant terms related to learning contexts that take place while learning how to use a product. These theories on learning build the foundation for the following literature analysis as they reflect the concerning paradigms that correspond to users when exposed to product learning situations (**Applied research approach**). Three major terms could be depicted which are then explored by a bibliometric literature analysis to determine their relevance in science and especially in the subject areas of marketing and information systems. Abstracts of the spotted publications in marketing and information systems research were analyzed to extract terms with high relevance for product-related customer learning. This first analysis is limited by the primarily determined learning paradigms, which may omit relevant keywords for further research. To overcome this eventuality a second analysis was conducted. Again a keyword-based literature approach was performed in chapter 3 to verify the relevance of these keywords according to the aspired research project as well as to identify other terms and research areas with high impact that could be excluded in the beginning due to the formerly selected learning-related keywords. Evaluation of results from the first attempt approved the selected terms but also additionally appropriate keywords were found. Finally chapter 4 gives a conclusion of the findings and also depicts limitations of the conducted research.

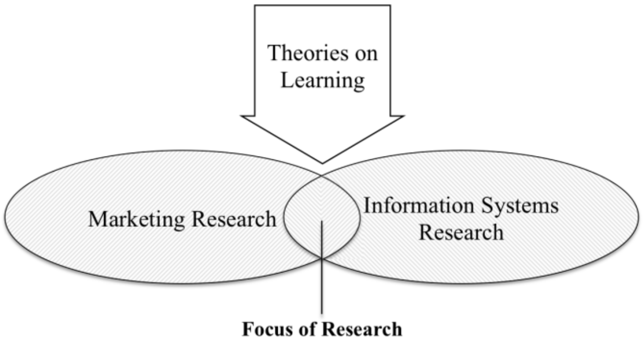


Fig.1. Applied research approach

II. LEARNING PARADIGMS IN RESEARCH

Various research fields such as Social Sciences, Psychology or Computer Sciences are related to learning theories but look at them from different perspectives. This chapter gives an insight into learning theory, which acts as an important foundation for further research in product and consumer learning. Product-related learning sets the focus on the learner and therefore requires adequate methods regarding style and format of learning to support knowledge building [12].

In literature various facets of learning are discussed. The Organization for Economic Co-operation and Development (OECD) defines three different forms of learning: formal, informal and non-formal learning [13-14]. While formal learning relates to an actively controlled process of knowledge transfer informal learning takes place without an intension of knowledge gaining. Learning is defined as non-formal when a formal learning environment enables learning in an informal way so an additional knowledge gain happens as a positive side-effect. OECD categorizes learning forms by learning awareness. Often mentioned dimensions of learning are self-regulated, self-directed and problem-based learning. These focus on the intention of learning and put personal learning habits into the center of attention. From a marketing perspective the respective intentions to learn how products work gives important insights for product positioning and feature implementation. This is highly related to personal learning habits, as product learning requires personalized actions to support a wide audience of potential customers. Research topics within the fields of self-regulated, self-directed and problem-based learning are expected to be applicable for further product learning research.

A. *Self-regulated Learning, Self-directed Learning, Problem-based Learning*

Learning situations in context of products are expected to take place when customers are not capable of performing a task to accomplish a goal. This is especially relevant for early product adoption phases where potential customers evaluate products by direct testing [15,16]. The customer identifies the gap between the current product related knowledge and an aspired outcome as a problem to overcome. This learning by doing approach puts customers in realistic usage situations where they encounter typical usage scenarios. Problem-based Learning is also well applied in pedagogic environments. Students get introduced to realistic cases, which enforce them to add additional meanings to the content of learning by applying problem-solving strategies. This matches with the problem situation to overcome when learning to use product features.

According to Loyens, Magda and Rikers [17] two main related fields of interest related to Problem-based Learning are Self-directed Learning and Self-regulated Learning. Self-regulated Learning describes learning as process where learners show active empathy and autonomy regarding their learning style and progress. Self-regulated learners are aware of their strengths and weaknesses and therefore react to these attitudes by individually set actions. Developmental, contextual, and individual boundaries and the self-motivated expansion of personal knowledge are the essential triggers for Self-regulated learners.

Self-directed Learning on the other hand sets a wider focus on proactive learning, which is often described in adult education or as life-long-learning [18]. It originates from the idea that learning does not happen in isolated environments but rather in an open interchange with others. The willingness to extend existing knowledge plays a major role though, as personal autonomy and self-managed learning processes have to take place. Smartphones or tablet computers with permanent connection to the Internet support these learning paradigms as they allow access to requested information independently from location or time.

B. *Research Method*

All these aspects view learning from different perspectives but characterize situations that have to be considered for gaining a deeper understanding on how product learning takes place. To examine already existing state-of-the-art literature on product learning and to also determine related research areas a literature analysis was conducted. The research profiling approach as a bibliometric method showed beneficial in similar research situations [19] and allows a widen analysis of existing publications. In contrast to classic literature reviews this approach gives profound insights into the topic by including a vast majority of relevant publications. For our purpose not the total number of publications is of importance as results will be used to determine keywords with highest relevance, rather than as a final source for further research.

The science database 'Scopus' as part of the 'SciVerse' platform was selected to perform this scientometric literature analysis as access to all publication abstracts was provided which showed beneficial for the second analysis. The research was conducted in February 2013. To give a wide overview of results no limitations but the search terms were set. In a first run the total available search results for the terms *problem-based learning*, *self-regulated learning* and *self-directed learning* within abstracts, keywords and titles of scientific publications were examined. After a first review the search terms were limited to keywords only as these are explicitly defined by the authors and therefore show more accurate results in terms of dealing with learning as key element of research.

C. *Results*

It can be stated that problem-based learning plays a major role in research in comparison to self-regulated and self-directed learning. While latters showed between 628 and 635 results problem-based learning outnumbers both with almost 85% of total related results in research.

Problem-based learning also shows a strong yearly growth in numbers of publications while only small increase in research related to self-regulated and self-directed learning can be spotted. More important than the number of publication are the subject areas where they got assigned to, as these show the fields of research with strong relevance for learning.

TABLE I. TOTAL AMOUNT OF SEARCH RESULTS ON SCIENCE DATABASE SCOPUS

	Self-regulated learning	**Self-directed learning**	**Problem-based learning**
Total results	1.170	1.959	8.338
Keyword results	628	635	6.796
2012	125	84	570
2011	125	102	607
2010	99	94	610
2009	72	78	653
2008	60	67	616
2007	30	47	532
2006	27	39	516
2005	14	25	434
2004	13	21	374
2003	11	18	352
2002	10	8	279

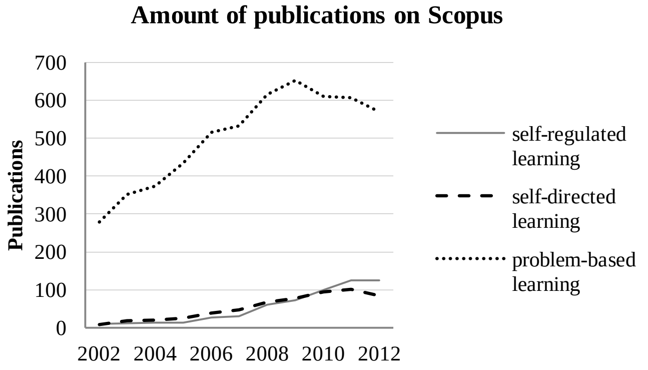


Fig.2. Search result growths on science database Scopus

TABLE II. SUBJECT AREAS OF PUBLICATIONS

Self-regulated learning	Self-directed learning	Problem-based learning
Social Sciences (391)	Social Sciences (294)	Medicine (3.429)
Psychology (222)	Medicine (192)	Social Sciences (2.661)
Computer Science (185)	Computer Science (154)	Nursing (1.582)
Engineering (47)	Engineering (93)	Engineering (618)
Mathematics (32)	Nursing (65)	Computer Science (471)
Arts and Humanities (28)	Psychology (30)	Biochemistry, Genetics and Molecular Biology (311)
Business, Management and Accounting (18)	Business, Management and Accounting (27)	Pharmacology, Toxicology and Pharmaceutics (254)
Medicine (15)	Pharmacology, Toxicology and Pharmaceutics (27)	Health Professions (201)
Nursing (6)	Mathematics (26)	Psychology (188)
Decision Sciences / Economics, Econometrics and Finance / Neuroscience (3)	Biochemistry, Genetics and Molecular Biology / Arts and Humanities (19)	Dentistry (150)

Profiling showed 418 publications as *undefined* subject areas for the search term *problem-based learning,* which were excluded from final results.

D. Discussion

Medical related fields of research such as Medicine in general (3.636 results), Nursing (1.653) or Dentistry (155) show strong interest in learning paradigms. Here the main focus is set on publications on the application of learning in working environments as these professions not only interact with but also attend other persons in a relationship between professional and patient. The second focus can be spotted in Social Science (3.336) and Psychology (440) as these areas relate to the behavior of students and teachers in learning environments.

TABLE III. MOST PUBLISHED RESEARCH AREAS

	Discipline	Self-regulated learning	Self-directed learning	Problem-based learning	Totals
1	Medicine	15	192	3.429	3.636
2	Social Sciences	391	294	2.661	3.346
3	Nursing	6	65	1.582	1.653
4	Computer Science	185	154	471	810
5	Engineering	47	93	618	758
6	Psychology	222	30	188	440
7	Biochemistry, Genetics and Molecular Biology	1	19	311	331
8	Pharmacology, Toxicology and Pharmaceutics	1	27	254	282
9	Health Professions	2	14	201	217
10	Dentistry	0	5	150	155

The gain in publications especially for problem-based learning shows an increasing interest for research in learning environments. While numbers in Social Sciences, Psychology, Engineering and Computer Science reflect psychological and technical aspects of learning, business related areas like marketing did occur far less often. The Scopus subject area *Business, Management and Accounting* lists 128 entries for the given search terms, which seems relatively under-represented, compared to very specific research areas such as Dentistry. Outcomes here should build a base for a deeper study of learning paradigms in general and should allow implications for product related learning strategies as well.

III. PRODUCT RELATED LEARNING IN RESEARCH

To identify the most represented keywords related to the application of learning in context of products or customers first study's search results were examined.

An abstract analysis determined *customer learning*, *consumer learning* as well as *product learning* as relevant terms for the application of learning approaches on product adoption and product support. Based on the outcomes of these results a second research profiling run was undertaken. Results of this second attempt should (1) list publications specifically relevant for product related learning (2) examine other potential keywords for further research and (3) ensure the focus in marketing and information systems for spotted publications to proof the relevance for these research areas.

A. Research Method

Again Scopus as science database was conducted for this second study. No limitations in terms of publication date, scientific field or publication type were given. The three key-phrases *customer learning, consumer learning* and *product learning* were concatenated by the *OR* parameter to find all relevant publications containing one or more of these terms in abstracts, titles, keywords and available full text papers as well.

A database-based information system was used to determine the amount of used keywords assigned by authors (author keywords) and the Scopus system (index keywords) [20]. Terms with different spelling in British and American English like *behavio(u)r* and *behavior* were merged to one result and numbers were added up when calculating result hits.

B. Results

Scopus allows the exportation of up to 2000 search results containing all descriptive meta-data except the full text. 1.811 results for the given query were returned which were then exported, manipulated and imported into a database system where each used keyword got stored as one row. All results were transformed to lower cases.

TABLE IV. TOTAL AND DISTINCT KEYWORDS IN SCOPUS

	Author keywords	**Index Keywords**
Total keywords	5.465	7.503
Distinct keywords	3.648	4.259
Uniqueness of keywords	66,75%	56,76%

Authors assigned 5.465 keywords in total with 3.648 distinct results while the Scopus index system assigned 7.503 in total with 4.259 distinct. The results highlight the dispersed contexts of product-related customer learning with more then 2/3rd unique used author keywords. Researchers entering this research are faced with these divergent results when attempting state-of-the-art research. The application of a scientometric literature analysis like the here conducted shows beneficial for research projects in similar interdisciplinary fields. While researchers attempt to assign their publications to at least one established and therefore strongly used keyword this cannot be stated for scientific publications on learning of product usage.

TABLE V. AUTHOR AND INDEX ASSIGEND KEYWORDS

Author keywords		**Index Keywords**		**Totals**	
Hits	**Keyword**	**Hits**	**Keyword**	**Hits**	**Keyword**
85	consumer behavio(u)r	63	marketing	99	consumer behavio(u)r
34	learning	63	article	88	marketing
32	advertising	60	learning systems	66	product development
29	brand equity	59	human	63	article
28	innovation	50	product development	62	decision making
27	internet	49	decision making	60	learning systems
25	marketing	38	female	59	human
22	consumer learning	38	mathematical models	58	customer satisfaction
21	customer satisfaction	37	customer satisfaction	56	learning
20	pricing	36	information systems	55	innovation
17	dynamic pricing	36	electronic commerce	53	electronic commerce
17	electronic commerce	35	adult	49	internet
17	brands	34	sales	45	advertising
17	new product development	33	male	38	sales
16	product development	31	neural networks	38	female
15	virtual worlds	29	costs	38	mathematical models
15	materialism	27	innovation	36	information systems
14	customer relationship management	26	adolescent	35	adult
14	e-commerce	26	project management	35	neural networks
14	knowledge management	25	product design	34	brand equity

Consumer behavio(u)r with a total of 85 hits is by far the most assigned keyword by authors with more generic keywords as *learning* (34) and *advertising* (32) following. Keywords seem to be dispensed as only these three have 30 hits or more. Index keywords on the other hand show higher repetition with 15 distinct keywords above 30 hits where 7 show more specific relation (learning systems, product development, decision making, customer satisfaction, information systems, electronic commerce, neural networks) to the field of research. In a last step all hits were again organized in one single list to show the total number of hits regardless if assigned by authors or Scopus.

C. Discussion

Results show essential for further research regarding product-related customer learning. It can clearly be stated that the topic is of high importance for the field of marketing. Beside the term *marketing* itself (88 hits), *consumer behavio(u)r* (99), *customer satisfaction* (58), *advertising* (45), *sales* (38) and *brand equity* (34) can be designated to the field of marketing. On the other hand also terms within the scientific discipline of information systems emerge such as *learning systems* (60), *internet* (49) and *information systems* (36) itself. Third, keywords are highly represented with a connection to marketing as well as information systems. Terms like *decision making* (62), *innovation* (55) or *electronic commerce* (53) belong to this third group, which also emphasizes the strong relation between both research areas in terms of product-related customer learning.

TABLE VI. MOST RELEVANT SEARCH TERMS AND CORRESPONDING RESEARCH AREAS

Marketing	**Information Systems**	**Marketing / Information Systems**
consumer behavio(u)r (99)	learning systems (60)	*decision making (62)*
marketing (88)	Internet (49)	innovation (55)
customer satisfaction (58)	information systems (36)	electronic commerce(53)
advertising (45)	-	-
sales (38)	-	-
brand equity (34)	-	-

By performing a content analysis for each single keyword an additional evaluation was processed. Except *decision making* where a strong focus on medical science areas can be spotted all determined keywords show high importance for the areas of information systems and marketing. While result of the first conducted research listed relevant areas of science this second attempt widened the spectrum by determining the relevant terms within these research areas. Both approaches start with different premises and so mutually complete the final results. The field of marketing, and especially consumer behavior and consumer satisfaction, shows high impact on learning of product functionalities and is thus of prior interest for further research.

IV. Conclusion and Limitations

Results point out that learning in context of product knowledge lists publications that mainly focus on the research fields of marketing and information systems. Of course the here-applied methodology approves less explorative for fields with clear assigned research. For wide spread areas such as product-related customer learning or other multidisciplinary fields this keyword-based approach shows the main involved scientific disciplines and the most relevant keywords for further research and is therefore beneficial in early research stages. This narrowing down of a very diverse field of research was conducted by a stepwise literature analysis. The first analysis represents the origins of research where a focus for a topic was set by the determination of relevant keywords for a bibliometric analysis. This analysis sets the base for the selection of relevant publications to determine required background knowledge before entering a research discipline. A second literature analysis based on keywords out of literature not only shows further meaningful keywords but also ensure the relevance of the selected keywords for specific research areas. Besides the three determined search terms *customer learning*, *consumer learning* and *product learning* new relevant terms within the disciplines of marketing and information systems could be depicted.

It should be stated that this research approach is also subject to some limitations. Although the popularity and dataset of the Scopus database by SciVerse, no other source in form of publication database was conducted. Even though for the purpose of this study the total amount of publications results is not of major relevance this limitation should be noted. Also the examined keywords were only revised in differences by spelling British and American English. No semantic relation between the terms was performed which therefore lists *customer* and *consumer* as different keywords although their similar meanings. At last also the assignment of keywords to disciplines of science is always nondistinctive as only the publication itself can directly be assigned to disciplines. For the purpose of this study the assignment was only performed to ensure a proper determined focus of research.

V. Further Research

The results determined in this paper lead to two main fields for further research. First, technical documentation for complex products as an important instrument of marketing has to be understood. Second, learning paradigms and theories are required to create beneficial product support for the customer. A mapping between both areas should demonstrate appropriate learning methods for different customer requirements due to new potentials in information systems. A strong focus hereby, as current results show, lies on increased customer satisfaction by applying mechanism to shorten learning efforts and building a support base for various forms of problem solving.

References

[1] Sueddeutsche Zeitung. http://www.sueddeutsche.de/leben/moderne-sammelwut-wenn-besitz-zur-last-wird-1.1089089, visited in March, 23, 2013.

[2] D. Gebert. Gebrauchsansweisungen als Marketinginstrument. Forkel-Verlag, Wiesbaden, 1988.

[3] TNS Infratest. Zukunft und Zukunftsfähigkeit der Informations- und Kommunikationstechnologien und Medien - Internationale Delphi-Studie 2030. 2009.

[4] J. Nielsen. Usability Engineering. Morgan Kaufmann, San Francisco, 1993.

[5] M. Reck. Internationale Kundenanforderungen an die Technische Dokumentation in Produktionsmaschinen. Schmidt-Römhild, Lübeck, 2008.

[6] J. Redish. "Technical Communication and Usability: Intertwined Strands and Mutual Influences Commentary", IEEE Transacions on Professional Communication, vol. 53, no. 3, pp. 191-201, 2010.

[7] M. Meuter, A. Ostrom, R. Roundtree, M. Bitner. "Self-Service Technologies: Understanding Customer Satisfaction with Technology-Based Service Encounters", Journal of Marketing, vol. 64, no. 3, pp. 50-64, 2000.

[8] M. Harnisch, T. Puchleitner, M. Reinisch, I. Uitz. "Model of a Personalization-based Agent System for Early Product Adoption Phases", Proceedings of 46th Hawaii International Conference on System Sciences, pp. 3455-3464, 2013.

[9] F. Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly, Society for Information Management and The Management Information Systems Research Center, vol. 13, no. 3, pp. 319-340, 1989.

[10] puls Marktforschung GmbH. Autokaeufer puls August 2012, Germany, 2012.

[11] Institute for Applied Ecology. PROSA Smartphones, Entwicklung der Vergabekriterien für ein klimaschutzbezogenes Umweltzeichen. Freiburg, August 2012.

[12] M. Scardamalia, C. Bereiter. "Knowledge Building", J. Guthrie (Ed.): Encyclopedia of Education, 2nd edition. New York, 2003.

[13] Organisation for Economic Co-operation and Development (OECD). Higher education and adult learning - Recognition of Non-formal and Informal Learning, http://www.oecd.org/edu/skills-beyond-school/recognitionofnon-formalandinformallearning-home.htm, visited March, 23, 2013.

[14] P. Werquin. Recognition of Non-Formal and Informal Learning: Country Practices, Organisation for Economic Co-operation and Development (OECD), 2010.

[15] G. Day, “The Product Life Cycle: Analysis and Applications Issues”, Journal of Marketing, vol. 45, no. 4, pp. 60-67, 1981.

[16] K.-P. Wiedmann, T. Frenzel, “Akzeptanz im E- Commerce – Begriff, Modell, Implikationen”, Konsumentenverhalten im Internet: Konzepte – Erfahrungen – Methoden, K.-P. Wiedmann, H. Buxel, T. Frenzel, and G. Walsh, Gabler, Wiesbaden, pp. 99-118, 2004.

[17] S. Loyens, J. Magda, R. Rikers. "Self-Directed Learning in Problem-Based Learning and its Relationships with Self-Regulated Learning", Educational Psychology Review, vol. 20, 2008.

[18] Organisation for Economic Co-operation and Development (OECD). OECD Observer Policy Brief: Lifelong Learning, February 2004.

[19] A. Porter, A. Kongthon, J. Lu. "Research Profiling: Improving the Literature Review", Scientom etrics, vol. 53, vo. 2, pp.351-370, 2002.

[20] Elsevier B.V. SciVerse Scopus - Content Coverage Guide, 2010

Detection and Isolation of Packet Dropping Attacker in MANETs

Ahmed Mohamed Abdalla, Ahmad H. Almazeed
Electronics Department, College of Technological Studies,
The Public Authority for Applied Education and Training,
P.O.Box 42325, Shuwaikh 70654, Kuwait

Imane Aly Saroit, Amira Kotb
Information Technology Department, Cairo University, 5
Dr. Ahmed Zewail St, Orman,
Giza 12613, Egypt

Abstract—**Several approaches have been proposed for Intrusion Detection Systems (IDS) in Mobile Ad hoc Networks (MANETs). Due to lack of MANETs infrastructure and well defined perimeter MANETs are susceptible to a variety of attacker types. To develop a strong security mechanism it is necessary to understand how malicious nodes can attack the MANETs. A new IDS mechanism is presented based on End-to-End connection for securing Optimized Link State Routing (OLSR) routing protocol. This new mechanism is named as Detection and Isolation Packet Dropped Attackers in MANETs (DIPDAM). DIPDAM mechanism based on three ID messages Path Validation Message (PVM) , Attacker Finder Message (AFM) and Attacker Isolation Message (AIM).**

DIPDAM mechanism based on End-to-End (E2E) communication between the source and the destination is proposed.

The simulation results showed that the proposed mechanism is able to detect any number of attackers while keeping a reasonably low overhead in terms of network traffic.

Keywords—MANETS; IDS; OLSR; DIPDAM

I. INTRODUCTION

A Mobile ad hoc Network (MANET) is a distributed and highly dynamic network environment. Mobility and unreliable wireless channels are the result of an unpredictable-dynamic network topology. Due to the fully distributed network, establishing a centralized node which can collect all of the network traffic is not feasible. In addition, mobile nodes have relatively limited power and bandwidth constraints, so they cannot carry high overhead security protection.

An ideal intrusion detection model in MANET should first have a reliable, distributed, low-overhead, message collecting, and exchanging mechanism. The mechanism should also adapt to changes in the network topology and tolerate message loss.

Second, the model should be affordable for low computation power devices. Third, the model should perform real-time protections since the routing topology may change very quickly and the attack damage may also propagate relatively quickly. Finally, the model should not generate high false positives and negatives with respect to new routing attacks.

The main goal in this paper is to detect successfully and isolate the data packet dropping attackers from routing path in OLSR routing protocol for MANETs.

In this paper, a new IDS mechanism is presented based on End-to-End connection for securing OLSR routing protocols. This new mechanism DIPDAM is based on three ID messages Path Validation Message (PVM) enables E2E feedback loop between the source and the destination, Attacker Finder Message (AFM) to detect attacker node through the routing path, and Attacker Isolation Message (AIM) to isolate the attacker from routing path and update the black list for each node then trigger to neighbors with updated information [1-2].

To save nodes resources, DIPDAM avoids monitoring every node at all times. DIPDAM is a fully distributed detection approach. DIPDAM is a scalable approach and allows the source to monitor its data messages with minimal overhead.

According to simulation results, It can be stated that DIPDAM mechanism can detect and isolate many types of misbehavior node(s) through the path between the source and the destination..

II. PREVIOUS WORK

Intrusion detection is defined as the method to identify “any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource”. [3]

Intrusion detection system (IDS) is a practical approach to enhance the security of existing networks. Briefly, an intrusion detection system monitors activity in a system or network in order to identify, to detect, and then to isolate current attacks.

There are three main components of an IDS:

- The collection of data.
- The analysis of collected data (Detection).
- The response of an alert when a threat is detected.

For Mobile Ad hoc Networks, the general function of an IDS is detecting misbehaviors by observing the networks traffic in a Mobile Ad hoc [4]. Most of recent researches focused on providing preventive schemes to secure routing in MANETs [5-9].

Key distribution and an establishment of a line of defense defined in [5], [6] based on mechanism in which nodes are either trusted or not and if trusted they are not compromised. Also contribution in [7], [9] considers the compromise of trusted nodes. It assumed a public key infrastructure (PKI) and a timestamp algorithm are in place. However, the above approaches cannot prevent attacks from a node who owns a legitimate key.

It is necessary to understand how malicious nodes can attack the MANETs. A model to address the Black Hole Search problem algorithm and the number of agents that are necessary to locate the black hole without the knowledge of incoming link Developed in [10]. Watchdog and path-rater are discussed in [11]. Their drawback is the increase of the percentage of overhead significantly with the percentage increase of misbehavior nodes. Ex-watchdog [12] suggests modifying the previous system to decrease the percentage of overhead. [13] Introduces IDS which formulate the problem of distributed collaborative defense against coordinated attacks as a dynamic game problem. The same group extends their work in [14] by proposing detection schemes that are suitable to detect in-band wormhole attacks. The first detection scheme uses the Sequential Probability Ratio Test (SPRT) is discussed in [15]. The SPRT has been proven to be an optimal detection test when the probability distributions of both normal and abnormal behaviors are given.

A feedback mechanism to secure OLSR against the link spoofing attacks was provided in [16], [17]. The solution assesses the integrity of control messages by correlating local routing data with additional feedback messages called CPM sent by the receivers of the control messages.

Another formal approach to harden the Multi Point Relay (MPR) selection and thwart the attacks against OLSR suggested in [18]. This approach validates the routing table and the topology information using trust based reasoning. Hence, each node can verify the validity of the received HELLO and TC messages simply by correlating the information provided by these messages. A technique to detect attacks by discussing a collusion attack model against the OLSR protocol was presented in [19].

III. Detection and Isolation of Packet Dropped Attackers in MANETs (DIPDAM)

New existing solutions for detecting data packet dropping in ad hoc networks work by monitoring individual nodes. Other solutions used so far for protecting these networks are authentication and encryption [20]. Most of these mechanisms are not considerably appropriate for MANETs resource constraints, i.e., bandwidth limitation and battery power, since they result in heavy traffic load for exchanging and verification of keys.

In DIPDAM mechanism, each source node in the network monitors its own packets (data packets or routing packets) by using a Path Validation Message (PVM) as shown in fig. 1. If a misbehavior node is detected, the other neighboring nodes are informed in order to help them in protecting themselves. Each source node monitors the behavior of its neighborhood instead of making each node in the networking doing this job which consumes nodes resources.

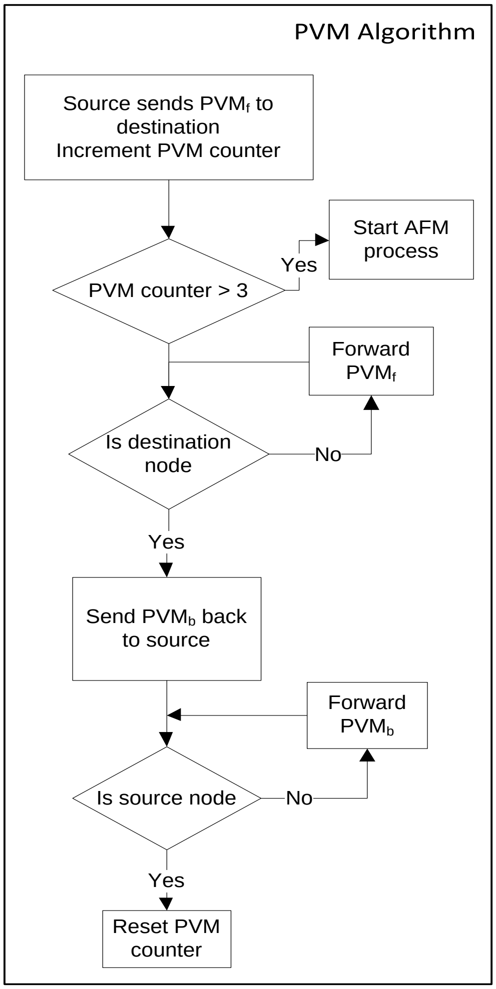


Fig.1. Flow chart for Path Validation Message (PVM) algorithm

A failure to get a reply for an N PVM messages sent (N is set to 3 in the flow chart), DIPDAM algorithm will trigger an Attacker Finder Message (AFM) algorithm shown in fig. 2.

The detector node needs to share the information about the detected attacker with other nodes in the network. This is accomplished by flooding the network with Attacker Isolation Messages (AIMs) [2]. It is noticed that nodes can be incorrectly detected as attackers due to network malfunction during a certain period. Such nodes would be wrongly isolated for the lifetime of the whole network.

A verification step is added to ensure that nodes are correctly detected and isolated. The process is illustrated in fig. 3. Fig. 4 shows a flow chart for the AIM algorithm.

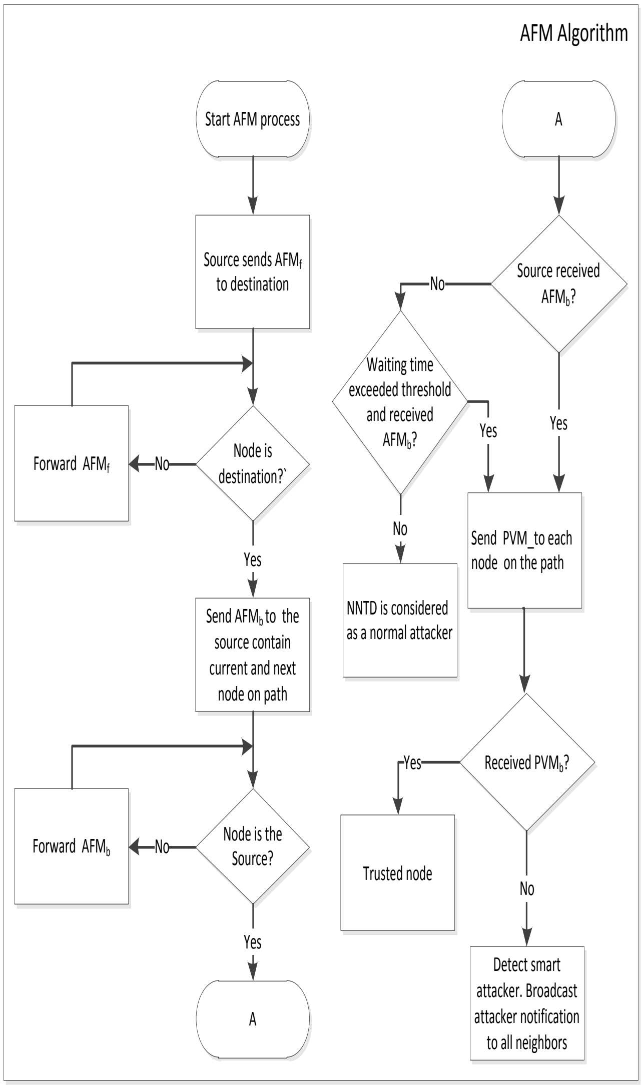

AFM Algorithm
Start AFM process
Source sends AFM$_f$ to destination
Node is destination?`
No
Forward AFM$_f$
Yes
Send AFM$_b$ to the source contain current and next node on path
Node is the Source?
No
Forward AFM$_b$
Yes
A
A
Source received AFM$_b$?
No
Waiting time exceeded threshold and received AFM$_b$?
No
NNTD is considered as a normal attacker
Yes
Yes
Send PVM_to each node on the path
Received PVM$_b$?
Yes
Trusted node
No
Detect smart attacker. Broadcast attacker notification to all neighbors

Fig.2. Flow chart for Attacker Finder Message (AFM) algorithm.

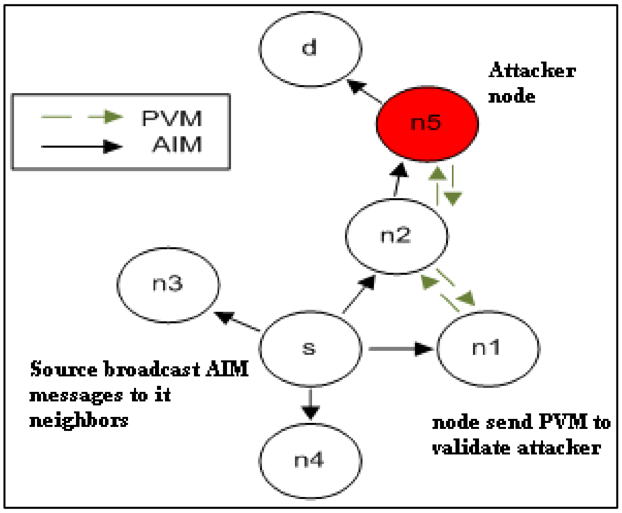

PVM
AIM
d
n5
Attacker node
n2
n3
s
n1
n4
Source broadcast AIM messages to it neighbors
node send PVM to validate attacker

Fig.3. Attacker Isolation Message (AIM) process.

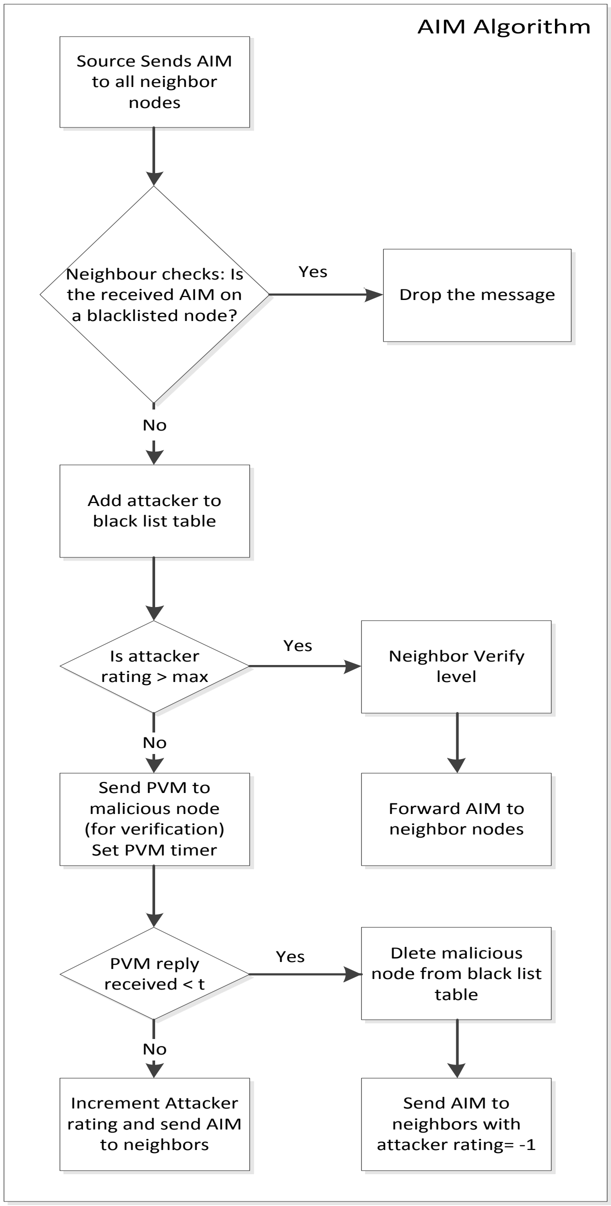

AIM Algorithm
Source Sends AIM to all neighbor nodes
Neighbour checks: Is the received AIM on a blacklisted node?
Yes
Drop the message
No
Add attacker to black list table
Is attacker rating > max
Yes
Neighbor Verify level
Forward AIM to neighbor nodes
No
Send PVM to malicious node (for verification) Set PVM timer
PVM reply received < t
Yes
Dlete malicious node from black list table
Send AIM to neighbors with attacker rating= -1
No
Increment Attacker rating and send AIM to neighbors

Fig.4. Flow chart for AIM algorithm.

To evaluate the robustness of DIPDAM mechanism we tested MANETs under different attacker types [21].

N1 nodes take contribution in the route discovery and route maintenance processes but refuses to forward data packets to protect its resources. This attack type can reduce network throughput, but does not affect any of the network traffic unless it is routed through selfish nodes, selfish nodes refuse to forward or drop data packets, this attacker type will be named as smart attacker.

N2 nodes neither contribute to the route discovery processes nor data-forwarding processes. Instead they use their resources only for transmissions of their own packets which are called selfish nodes. An attacker with this criterion will be named normal attacker.

N3 nodes behave properly if its energy level lies between full energy-level and certain threshold T1. They behave like node of type N2 if energy level lies between threshold T1 and another threshold T2 and if energy level falls below T2, they behave like node of type N1.

N1, N2, and N3 nodes are risky to routing protocols. These nodes suspend the data flow by either dropping or refusing to forward the data packets thus forcing routing protocol to select an alternative available route which it may again contain some malicious nodes, resulting in the new route also to fail. This process form a loop which enforce source to conclude that data cannot be further transferred.

The proposed work is designed to detect and isolate N1 type and N2 type. N3 type selfish nodes will be detected only when they behave similar to N1 or N2 type nodes.

Dropping any packets affects the network performance by causing the retransmission of data packets many times. Furthermore, it can prevent the end-to-end communications between nodes.

IV. NETWORK SIMULATOR PROGRAM

The NS-2 simulation tool [22-23] consists of two sets of scenario; topology scenario and traffic generation pattern. The topology scenario defines the simulation area and the mobility model of randomly distributed mobile nodes over the simulation time. The traffic pattern defines the characteristics of data communications, data packet size, packet type, packet transmission rate and number of traffic flows. Each node is assumed to be equipped with a wireless transceiver operating on 802.11 wireless standards. The physical radio frequency characteristics of each wireless transceiver such as transmit power, the antenna gain, and signal to noise and interference ratio, are chosen with a bit rate of 2Mb/sec and a transmission range of 250 meters with an omni-directional antenna.

The simulation scenarios consist of two different settings. First, the impact of network density or size is assessed by varying the number of mobile nodes placed on an area of a fixed size of 1500m x 300m. The second simulation scenario investigates the effects of node mobility on the performance of route discovery by varying the maximum speed of mobile nodes placed on a fixed area of 1500m x 300m.

Each node participating in the network is transmitting within the 250m transmission range, and each simulation runs for a period of 900sec. The above settings could represent a MANET scenario in real life; like a University campus. Note that the number of mobile nodes could be larger than the one presented in these scenarios and the operational time could be longer; the values chosen are to keep the simulation running time manageable while still generating enough traces for analysis. Flows of Constant Bit Rate (CBR) unicast data packets, each with size 512 bytes.

In this study, mobile nodes move according to the widely used random waypoint mobility model where each node at the beginning of the simulation remains stationary for pause time seconds, then chooses a random destination and starts moving towards it with a speed selected from a uniform distribution [0, V max]. Other simulation parameters used in this research study have been widely adopted in existing performance evaluation studies of MANETs and are summarized below in Table 1.

TABLE I. SYSTEM PARAMETERS USED IN THE SIMULATION EXPERIMENTS.

Simulation Parameter	Value
Simulator	NS-2 (v.2.31)
Transmitter range	250 meter
Bandwidth	2 Mbps
Traffic type	CBR
Number of Nodes	30
Topology size	1500m x 300m
Packet size	512 bytes
Simulation time	900 sec

V. PERFORMANCE METRICS

In order to evaluate the performance of our proposed Intrusion Detection System DIPDAM, we will focus mainly on evaluating four performance metrics:-

a) *Average overhead:*

The average overhead is defined as the total number of data packet and routing control packets normalized by the total number of received data packets.

b) *Average Packet Delivery Ratio (Rating):*

It is the ratio of the number of packets received successfully to the total number of packets transmitted.

c) *Average Packet dropping:*

The average packet dropping is the average percentage of data packet dropped to all data and control packets sent from the sources to the destinations.

d) *Average end-to-end delay:*

The end-to-end-delay is the average overall delay measured from the sources to the destinations.

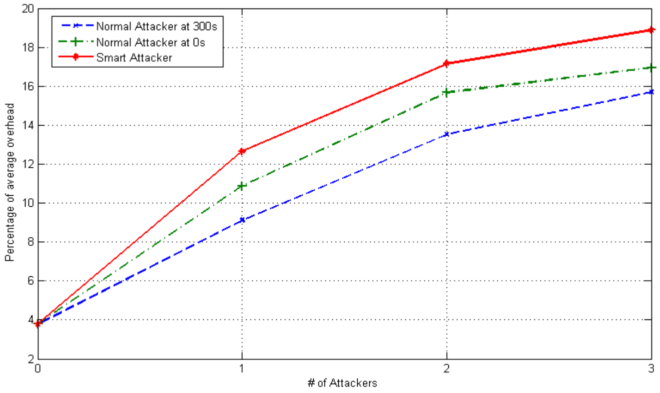


Fig.5. Average percentage of overhead vs. numbers of attackers.

Figure 5 shows that the average overhead increases directly with the numbers of attackers. The increase in the percentage of overhead compared to the original OLSR came from three major reasons. Firstly, PVM messages inserted within data packet to monitor the path between the source and the destination. Secondly, due to AFM messages used to find attackers through the transmission path. Finally, because of AIM messages needed to isolate the attacker from the routing path.

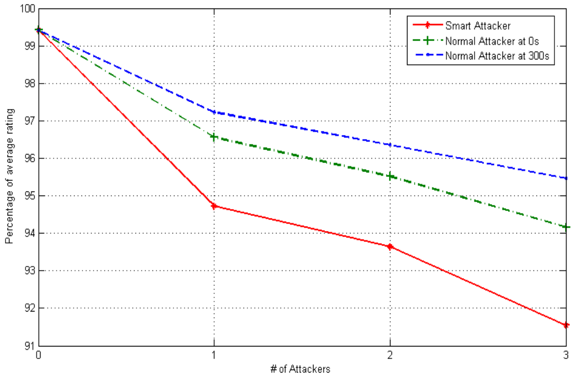


Fig.6. Percentage of average rating vs. numbers of attackers.

As shown from figure 6 the percentage of average rating almost decreases linearly with the increase of the number of attackers. The decrease is due to the dropped data attacker found in routing path.

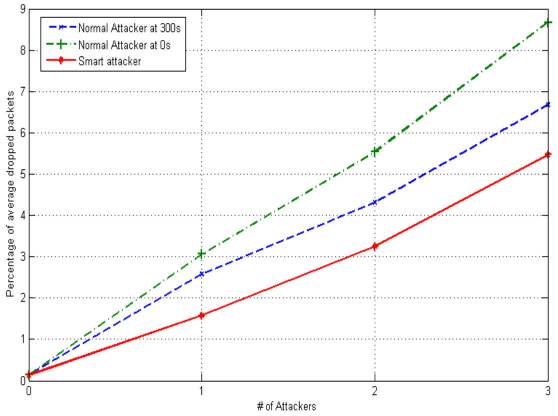


Fig.7. Percentage of average dropped packets vs. number of attackers.

As shown in figure 7, the percentage of average dropped packets almost increases linearly with the increase of the number of attackers. The increase in is due to the dropped data attacker found in routing path.

Average End-to-End delay versus the number of attackers is shown in fig. 8.

Results obtained in the above figure illustrate an increase in the average delay as the number if attackers increase. The increase of E2E delay comes from two major reasons.

Firstly, the network takes some time to detect and isolate the attacker. Secondly, since the attacker damaged the routing path, the process to recalculate an alternative routing path needs extra time which results in the increase of the average E2E delay time

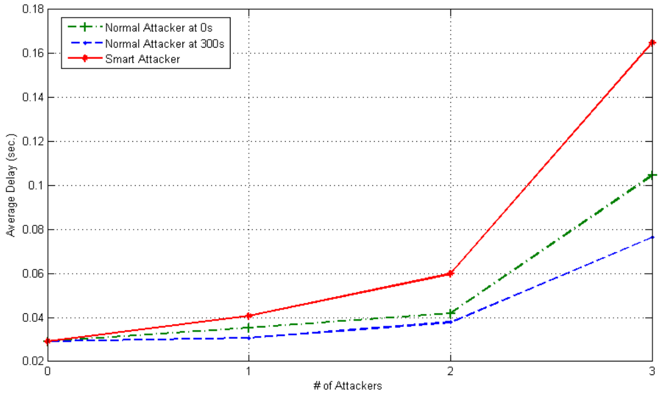


Fig.8. Average End to End Delay vs. number of attackers.

Table 2 shows a sample from the detection process results. The table contains the attacker detector node, transmission path, attacker node, and the attacker type. Table 2 shows that source nodes 13, 19, and 28 were able to detect the attacker's nodes 12, 14, and 21 successfully when present in the path. The results show that the accuracy of detection is independent of the path length or the location of the attacker in respect to the detector. The detection procedure also can detect different types of attackers found in the network at the same time

TABLE II. SAMPLE LIST OF ATTACKERS DETECTED.

Detector	Path	Attacker	Attacker type
28	28.12.23.26.26.23.0	12	Normal
28	28.1.14.0	14	Smart
28	28.21.5.0	21	Normal
19	19.16.9.12.0	12	Normal
19	19.7.14.23.0	14	Smart
19	19.21.13.0	21	Normal
13	13.21.4.0	21	Normal

VI. DISCUSSION

From the above figures, It can be concluded that DIPDAM mechanism achieved better performance metrics when the attacker is a normal attacker and its attacking action after certain amount of time from the beginning of the simulation test.

On the other hand the smart attacker type take larger time, higher overload, more dropping packet, and worst average rating compared to other attacker types discussed. It is expected that this result is due to deep processing to detect and isolate the smart attackers.

VII. CONCLUSION

We have presented IDS mechanism based on End-to-End connection for securing OLSR routing protocol. DIPDAM mechanism can detect and isolate many types of misbehavior node(s) through the path between the source and the destination then a blacklist of misbehavior nodes is created and broadcasted to 1-Neighbors IDS mechanism was proposed for Detection and Isolation of Packet-Dropped Attacker in MANETs (DIPDAM).

DIPDAM, a fully-distributed message exchange framework designed to overcome the challenges caused by the decentralized and dynamic characteristics of MANETs.

DIPDAM performance was inspected using different comparable performance metrics to show its reliability and efficiency in detection and isolation many types of misbehavior nodes.

Three ID messages are proposed to implement DIPDAM Path Validation Message (PVM) enables E2E feedback loop between the source and the destination, Attacker Finder Message (AFM) to detect attacker node through the routing path, and Attacker Isolation Message (AIM) to isolate the attacker from routing path and update the black list for each node then trigger to neighbors with updated information.

VIII. Future work

Our mechanism must be tested in real MANETs with different conditions like variation on mobility, size, network traffic type, and node density.

DIPDAM mechanism may be upgraded to detect both types of attackers, data packet attackers and route packets attackers.

The same mechanism can be tried on different Manet’s protocols from other categories.

References

[1] Ahmed M. Abdalla, Imane A. Saroit, Amira Kotb, Ali H. Afsari,"An IDS for Detecting Misbehavior Nodes in Optimized Link State Routing Protocol", International Journal of Advanced Computer Science, Vol. 1, No. 2, Pp. 87-91, Aug. 2011.

[2] Ahmed M. Abdalla, Imane A. Saroit, Amira Kotb, Ali H. Afsari," Misbehavior Nodes Detection and Isolation for MANETs OLSR Protocol", World Conference on Information Technology. Procedia Computer Science volume 3, 2011, pages 115–121.

[3] Y. Huang and Wenke Lee, "Attack analysis and detection for ad hoc routing protocols", In Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID'04), pages 125-145. Springer, 2004.

[4] A. Fourati, K. Al Aghha "An IDS First Line of defense for Ad Hoc Networks", in Proceeding of 2007 IEEE WCNC.

[5] Y-C. Hu, A. Perrig, and D. B. Johnson, “Ariadne: A secure On-Demand Routing Protocol for Ad hoc Networks,” in Proceedings of the MobiCom 2002, Atlanta, Georgia, USA, September 23-28, 2002.

[6] C. Adjih, Th. Clausen, Ph. Jacquet, A. Laouiti, P. Muhlethaler, and D. Raffo, “Securing the OLSR protocol,” In Proceedings of Med-Hoc-Net, Mahdia, Tunisia, June 25, 2003.

[7] D. Dhillon, T.S. Randhawa, M. Wang and L. Lamont, “Implementing a Fully Distributed Certificate Authority in an OLSR MANET,” IEEE WCNC2004, Atlanta, Georgia USA, March 21-25, 2004.

[8] D. Raffo, C. Adjih, T. Clausen, and P. Muhlethaler, “An Advanced Signature System for OLSR,” in Proceedings of the 2004 ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN 04), Washington, DC, USA, October 25 2004.

[9] C. Adjih, D. Raffo, and P. Muhlethaler, “Attacks Against OLSR: Distributed Key Management for Security,” 2nd OLSR Interop/ Workshop, Palaiseau, France, July 28-29, 2005.

[10] [Peter Glaus, "Locating a Black Hole without the Knowledge of Incoming Link", Algorithmic Aspects of Wireless Sensor Networks, Lecture Notes in Computer Science, Volume 5304. Springer-Verlag Berlin Heidelberg, 2009, p. 128, http://www.springerlink.com/index/h8424573040077v5.pdf

[11] S. Marti, T. J. Giuli, K. Lai, and M. Baker, "Mitigating Routing Misbehavior in Mobile Ad hoc Network", in 6th International Conference on Mobile Computing and Networking, MOBICOM'00, p255-265, Aug 2000.

[12] Nidal Nasser, Yunfeng chen, "Enhanced Intrusion Detection System for Discovering Malicious Node in Mobile Ad hoc Networks", Communications, 2007. ICC '07. IEEE International Conference on Publication Date: 24-28 June 2007, page(s): 1154-1159.

[13] Baras, John S. Radosavac, Svetlana Theodorakopoulos, George Sterne, Dan Budulas, Peter Gopaul, Richard " Intrusion Detection System Resiliency to Byzantine Attacks: The Case Study of Wormholes in OLSR", Military Communications Conference, 2007. MILCOM 2007. IEEE Publication Date: 29-31 Oct. 2007, page(s): 1-7, Orlando, FL, USA.

[14] Shanshan Zheng Tao Jiang Baras, J.S. Sonalker, A. Sterne, D. Gopaul, R. Hardy, R. "Intrusion detection of in-band wormholes in MANETs using advanced statistical methods", Military Communications Conference, 2008. MILCOM 2008. IEEE Publication Date: 16-19 Nov. 2008, page(s): 1-7, San Diego, CA.

[15] M.T. Refaei, Yanxia Rong, L. A. DaSilva, and Hyeong-Ah Choi, "Detecting Node Misbehavior in Ad hoc Networks", Communications, 2007. ICC '07. IEEE International Conference on Publication Date: 24-28 June 2007, page(s): 3425-3430, Glasgow.

[16] J.P. Vilela and J. Barros, "A Feed Reputation Mechanism to Secure the Optimized Link State Routing Protocol", The 3rd IEEE/CreateNet International Conference on Security and Privacy in Communication Networks, Nice, France, September 2007.

[17] J.P. Vilela and J. Barros, "A Cooperative Security Scheme for Optimized Link State Routing in Mobile Ad-hoc Networks", Proc of the 15th IST Mobile and Wireless Communications Summit, Mykonos, Greece, June 2006.

[18] Asmaa Adnane , Rafael T. de Sousa, Jr., Christophe Bidan, Ludovic Mé, " Autonomic trust reasoning enables misbehavior detection in OLSR", Proceedings of the 2008 ACM symposium on Applied computing, Pages 2006-2013.

[19] B. Kannhavong, H. Nakayama, N. Kato, Y. Nemoto, and A. Jamalipour, "A Collusion Attack Against OLSR-based Mobile Ad Hoc Networks", in Proceeding of 2006 IEEE GLOBECOM.

[20] Y. Rebahi, V. Mujica, C. Simons, and D. Sisalem, "SAFE: Securing pAcket Forwarding in ad hoc nEtworks", In 5th Workshop on Applications and Services in Wireless Networks 2005.

[21] Sevil Sen, "Evolutionary Computation Techniques for Intrusion Detection in Mobile Ad Hoc Networks", PhD Thesis, University of York Department of Computer Science, March 2010.

[22] The Vint Project, "The Network Simulator –ns-2," http://www.isi.edu/nsnam/ns/index.html

[23] F. J. Ro, "UM-OLSR Documentation," University of Murcia, March 2005, http://masimum.dif.um.es/um-olsr/html

Comparative Analysis of K-Means and Fuzzy C-Means Algorithms

Soumi Ghosh
Department of Computer Science and Engineering,
Amity University, Uttar Pradesh
Noida, India

Sanjay Kumar Dubey
Department of Computer Science and Engineering,
Amity University, Uttar Pradesh
Noida, India

***Abstract*—In the arena of software, data mining technology has been considered as useful means for identifying patterns and trends of large volume of data. This approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. It is a computational intelligence discipline which has emerged as a valuable tool for data analysis, new knowledge discovery and autonomous decision making. The raw, unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the assignment of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. The outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. There are various algorithms which are used to solve this problem. In this research work two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms are compared. These algorithms are applied and performance is evaluated on the basis of the efficiency of clustering output. The numbers of data points as well as the number of clusters are the factors upon which the behaviour patterns of both the algorithms are analyzed. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means clustering.**

Keywords—clustering; k-means; fuzzy c-means; time complexity

I. INTRODUCTION

In the field of software data analysis is considered as a very useful and important tool as the task of processing large volume of data is rather tough and it has accelerated the interest of application of such analysis. To be precise data mining is the analysis of datasets that are observational, aiming at finding out unsuspected relationships among datasets and summarizing the data in such a noble fashion that are both understandable and useful to the data users [9].

It also makes data description possible by means of clustering visualization, association and sequential analysis. Data clustering is primarily a method of data description which is used as a common technique for data analysis in various fields like machine learning, data mining, pattern recognization, image analysis and bio-informatics. Cluster analysis is also recognised as an important technique for classifying data, finding clusters of a dataset based on similarities in the same cluster and dissimilarities between different clusters [13]. Putting each point of the dataset to exactly one cluster is the basic of the conventional clustering method where as clustering algorithm actually partitions unlabeled set of data into different groups according to the similarity. As compare to data classification, data clustering is considered as an unsupervised learning process which does not require any labelled dataset as training data and the performance of data clustering algorithm is generally considered as much poorer. Although data classification is better performance oriented but it requires a labelled dataset as training data and practically classification of labelled data is generally very difficult as well as expensive. As such there are many algorithms that are proposed to improve the clustering performance. Clustering is basically considered as classification of similar objects or in other words, it is precisely partitioning of datasets into clusters so that data in each cluster shares some common trait. The hierarchical, partitioning and mixture model methods are the three major types of clustering processes that are applied for organising data. The choice of application of a particular method generally depends on the type of output desired, the known performance of the method with particular type of data, available hardware and software facilities and size of the dataset [13].

In this research paper, K-Means and Fuzzy C-Means clustering algorithms are analyzed based on their clustering efficiency.

II. K-MEANS CLUSTERING

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters.

Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution [2, 6, 14]. In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose coordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

Algorithmic steps for K-Means clustering [12]

1) Set K – To choose a number of desired clusters, K.

2) Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.

3) Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.

4) Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.

5) Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move.

The actual data samples are to be collected before the application of the clustering algorithm. Priority has to be given to the features that describe each data sample in the database [3, 10]. The values of these features make up a feature vector ($F_{i1}, F_{i2}, F_{i3,\ldots\ldots\ldots..,}$ F_{im}) where F_{im} is the value of the M-dimensional space [12]. As in the other clustering algorithms, k- means requires that a distance metric between points is to be defined. This distance metric is used in the above mentioned step (iii) of the algorithm. A common distance metric is the Euclidean distance. In case, the different features used in the feature vector have different relative values and ranges then the distance computation may be distorted and so may be scaled.

The input parameters of the clustering algorithm are the number of clusters that are to be found along with the initial starting point values. When the initial starting values are given, the distance from each sample data point to each initial starting value is found using equation. Then each data point is placed in the cluster associated with the nearest starting point. After all the data points are assigned to a cluster, the new cluster centroids are calculated. For each factor in each cluster, the new centroid value is then calculated. The new centroids are then considered as the new initial starting values and steps (iii) and (iv) of the algorithm are repeated. This process continues until no more data point changes or until the centroids no longer move.

III. Fuzzy C-Means Clustering

Bezdek [5] introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition [16].

With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy clustering theory was proposed in 1980's. This algorithm is used for analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster.

Infact, FCM is a data clustering technique [11, 7] in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the center of a cluster which will have a low degree of belonging to that cluster.

Algorithmic steps for Fuzzy C-Means clustering [13]

We are to fix c where c is (2<=c<n) and then select a value for parameter 'm' and there after initialize the partition matrix $U^{(0)}$. Each step in this algorithm will be labelled as 'r' where r = 0, 1, 2 …

1) We are to calculate the c center vector {V_{ij}} for each step.

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ij})^m} \qquad (1)$$

2) Calculate the distance matrix $D_{[c,n].}$

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2} \qquad (2)$$

3) Update the partition matrix for the r^{th} step, $U^{(R)}$ as

$$\mu_{ij}^{r-1} = \left(1 \Big/ \sum_{j=1}^{c} (d_{ik}^{r} \;/ d_{jk}^{r}) \;^{2/m-1} \right) \qquad (3)$$

If $\|U^{(k+1)}-U^{(k)}\|<\delta$ then we are to stop otherwise we have to return to step 2 by updating the cluster centers iteratively and also the membership grades for data point [13].

FCM iteratively moves the cluster centers to the right location within a dataset. To be specific introducing the fuzzy logic in K-Means clustering algorithm is the Fuzzy C-Means algorithm in general. Infact, FCM clustering techniques are based on fuzzy behaviour and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all. This algorithm is basically similar in structure to K-Means algorithm and it also behaves in a similar fashion.

IV. Implementation Methodology

For the purpose of testing the efficiency of K-Means and FCM in matlab [8], the well known UCI Machine Learning Repository [1] is used and it is actually a collection of databases which is widely used by the researchers of Machine Learning, especially for the empirical algorithms analysis of this discipline [1]. Iris plant Dataset: Total number of attributes is five of which four (Sepal Length, Sepal Width, Petal Length and Petal Width) are numeric and one is non-numeric. This non-numeric attribute has three classes. The total numbers of instances are 150 in this attribute. The three classes are Iris Setosa, Iris Versicolour, and Iris Virginica. One class is linearly separable from the other 2, the latter are not linearly separable from each other.

A. Implementation of K-Means Clustering

The matlab function kmeans used for K-Means clustering to partitions the points in the n-by-p data matrix data into k clusters [8]. This iterative partitioning minimises the overall sum of clusters, within cluster sums of point- to cluster centroid distances. Rows of data correspond to points, columns correspond to variables and kmeans return an n-by-1 vector idx containing the cluster indices of each point. By default, k-means uses squared Euclidean distances. When data is a vector, k-means treats it as an n-by-1 data matrix, regardless of its orientation. The iris dataset for three clusters, five 'replicates' have been specified and the 'display' parameters are used to print out the final sum of distances for each of the solutions. The sum total of distances covering 13 iterations that have taken into considerations in this paper comes to 7897.88. The total **elapsed time is 0.443755 seconds.** Following scattered K-Means graph for iris data set (sepal length, sepal width and petal length) represents three clusters.

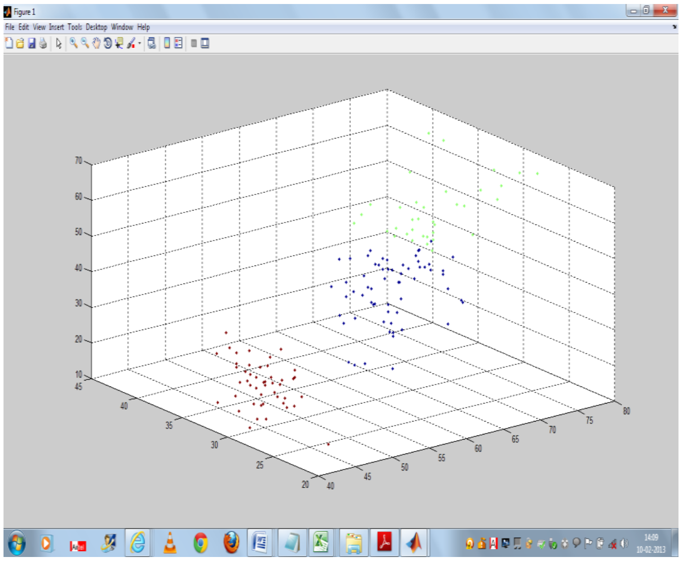


Fig.1. Scattered K-Means graph of iris dataset for three clusters

B. Implementation of Fuzzy C-Means Clustering

The mat lab function fcm performs FCM clustering [8]. The function fcm takes a data set and a desired number of clusters and returns optimal cluster centers and membership grades for each data point. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Next, fcm assigns every data point a membership grade for each cluster.

By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade. The dataset is obtained from the data file 'iris.dat'[1]. From each of the three groups (setosa, versicolor and virginica), two characteristics (for example, sepal length vs. sepal width) of the flowers are plotted in a 2-dimensional plot.

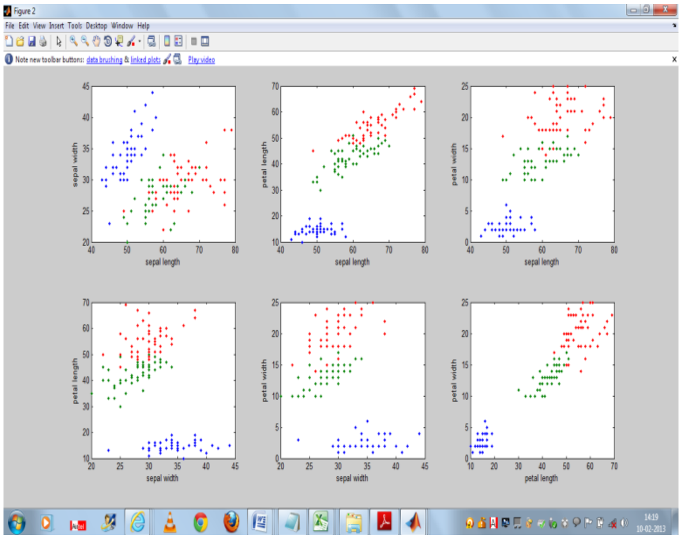

Note new toolbar buttons: data brushing & linked plots Play video

Fig.2. Scattered Fuzzy C-Means graph of iris dataset for three clusters

FCM clustering is an iterative process. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. For iris dataset comprising of 30 total iteration count results a total objective function equals to 6058.689983. The total **elapsed time is 0.781679 seconds.** The figure shows the initial and final fuzzy cluster centers. The bold numbers represent the final fuzzy cluster centers obtained by updating them iteratively.

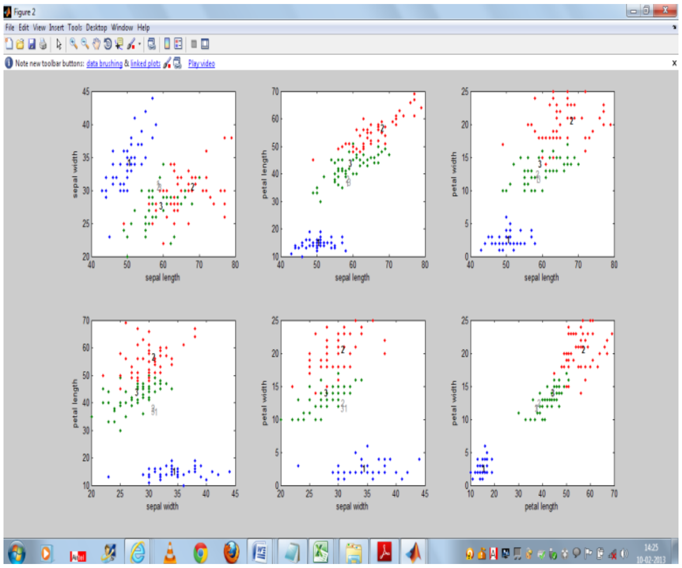

Note new toolbar buttons: data brushing & linked plots Play video

Fig.3. Scattered Fuzzy C-Means graph with initial and final fuzzy cluster centers

V. Experimental Results

This experiment reveals the fact that K-Means clustering algorithm consumes less elapsed time i.e. 0.443755 seconds than FCM clustering algorithm which takes 0.781679 seconds. On the basis of the result drawn by this experiment it may be safely stated that K-Means clustering algorithm less time consuming than FCM algorithm and hence superior.

A. *Comparison of Time Complexity of K-Means and FCM*

The time complexity of K-means [15] is O(ncdi) and time complexity of FCM [4] is O(ndc^2i). Keeping the number of data points constant we may assume that n = 100, d = 3, i = 20 and varying number of clusters where n = number of data points, c = number of cluster, d = number of dimension and i = number of iterations. The following table and graph represents the comparison in details.

TABLE I. COMPARATIVE ANALYSIS OF K-MEANS AND FCM

Algorithm	**Time Complexity**	**Elapsed Time (Seconds)**
K-Means	O(ncdi)	0.443755
FCM	O(ndc^2i)	0.781679

TABLE II. TIME COMPLEXITY OF K-MEANS AND FCM WHEN NUMBER OF CLUSTERS VARYING

S.No.	**Number of Clusters**	**K-Means Time Complexity**	**FCM Time Complexity**
1	1	6000	6000
2	2	12000	24000
3	3	18000	54000
4	4	24000	96000

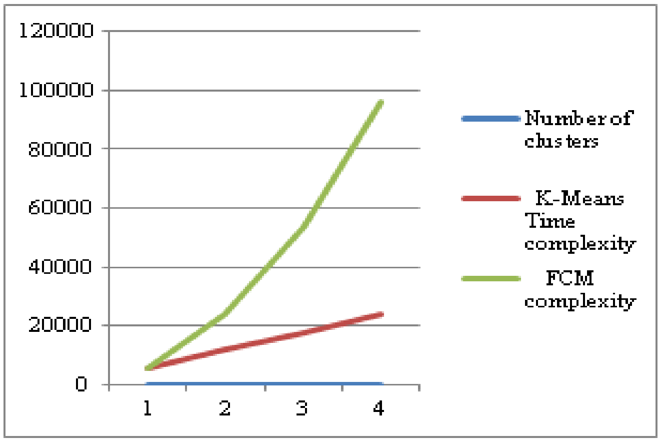


Fig.4. Time complexity of K-Means and FCM by varying number of clusters

Now keeping no. of cluster constant, lets assume n=150, d=2, c=2 and varying no. of iteration, we obtain the following table and graph.

TABLE III. TIME COMPLEXITY OF K-MEANS AND FCM WHEN NUMBER OF ITERATIONS VARYING

S.No.	**Number of Iterations**	**K-Means Time Complexity**	**FCM Time Complexity**
1	5	3000	6000
2	10	6000	12000
3	15	9000	18000
4	20	12000	24000

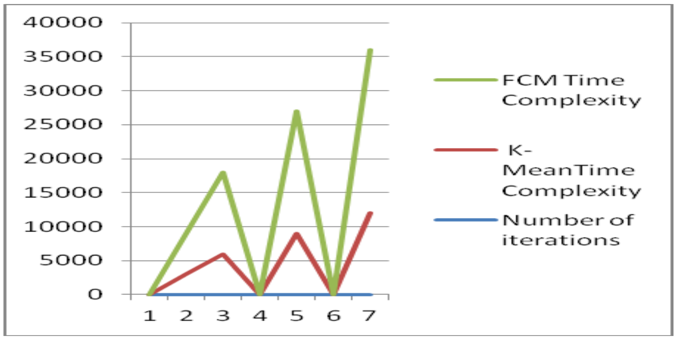


Fig.5. Time complexity of K-Means and FCM by varying number of iterations

VI. CONCLUSION

K-Means partitioning based clustering algorithm required to define the number of final cluster (k) beforehand. Such algorithms are also having problems like susceptibility to local optima, sensitivity to outliers, memory space and unknown number of iteration steps that are required to cluster. The time complexity of the K-Means algorithm is O(ncdi) and the time complexity of FCM algorithm is O(ndc^2i). From the obtained results we may conclude that K-Means algorithm is better than FCM algorithm. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. Infact, FCM clustering which constitute the oldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. They have been mainly used for discovering association rules and functional dependencies as well as image retrieval. So, overall conclusion is that K-Means algorithm seems to be superior than Fuzzy C-Means algorithm.

REFERENCES

[1] A. Asuncion and D. J. Newman, UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science, 2013.

[2] A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A review", ACM Computing Surveys, vol. 31, no. 3, 1999.

[3] A. Rakhlin and A. Caponnetto, "Stability of K-Means clustering", Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2007, pp. 216–222.

[4] A. Rui and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for Classification", International Symposium on Evolving Fuzzy Systems, 2006 , pp. 112-117.

[5] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum Press, 1981.

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2nd ed., New Delhi, 2006.

[7] L. Hui, "Method of image segmentation on high-resolution image and classification for land covers", Fourth International Conference on Natural Computation, vol. 5, 2008, pp. 563-566.

[8] Mathworks. http: //www.mathworks.com

[9] R. Mosley, "The Use of Predictive Modeling in the Insurance Industry", Pinnacle actuarial resources, 2005.

[10] S. Borah and M. K. Ghose, "Performance analysis of AIM-K-Means and K-Means in quality cluster generation", Journal of Computing, vol. 1, Issue-1, 2009.

[11] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure", IEEE Transactions on Systems, Man and Cybernetics, vol. 34, 1998, pp. 1907-1916,.

[12] T. Kanungo and D. M. Mount, "An Efficient K-means Clustering Algorithm: Analysis and Implementatio*n",* Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 24, no. 7, 2002.

[13] V. S. Rao and Dr. S. Vidyavathi, "Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on Iris data", Indian Journal of Computer Science and Engineering, vol.1, no.2, 2010 pp. 145-151.

[14] X. Hui, J. Wu and C. Jian, "K-Means clustering versus validation measures: A data distribution perspective", IEEE Transactions on Systems, Man, and cybernetics, vol. 39, Issue-2, 2009 , pp.319-331.

[15] X. Rui, D. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, vol.16, no.3, 2005.

[16] Y. Yong, Z. Chongxun and L. Pan, "A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding", Measurement Science Review, vol. 4, no.1, 2004.

AUTHORS PROFILE

Soumi Ghosh is pursuing M. Tech (CS&E) at Amity University, Uttar Pradesh, India. Her research areas include Software Engineering and Fuzzy Logic.

Mr. Sanjay Kumar Dubey is Assistant Professor and Proctor in Amity University, Uttar Pradesh, India. He has submitted his Ph. D. thesis in Object Oriented Software Engineering. He has published more than 73 papers in International Journals. He has presented 14 research papers at various National/International conferences.He is member of IET and IEANG. His research areas include Human Computer Interaction, Soft Computing and Usability Engineering.

Developing a Stochastic Input Oriented Data Envelopment Analysis (SIODEA) Model

Basma E. El-Demerdash
Teaching Assistant,
Department of Operations Research and Decision Support,
Faculty of Computers and Information, Cairo University,
Egypt

Ihab A. El-Khodary
Associate Professor,
Department of Operations Researchand Decision Support,
Faculty of Computers and Information, Cairo University,
Egypt

Assem A. Tharwat
Professor, Dean of the Higher
Canadian International College for
Engineering and Business,
Canadian International College,
Egypt

Abstract—**Data Envelopment Analysis (DEA) is a powerful quantitative tool that provides a means to obtain useful information about efficiency and performance of firms, organizations, and all sorts of functionally similar, relatively autonomous operating units, known as Decision Making Units (DMU). Usually the investigated DMUs are characterized by a vector of multiple inputs and multiple outputs. Unfortunately, not all inputs and/or outputs are deterministic; some could be stochastic. The main concern in this paper is to develop an algorithm to help any organization for evaluating their performance given that some inputs are stochastic. The developed algorithm is for a Stochastic Input Oriented Model based on the Chance Constrained Programming, where the stochastic inputs are normally distributed, while the remaining inputs and all outputs are deterministic.**

Keywords—Data Envelopment Analysis; Stochastic Variables; Input Oriented; Performance Measure; Efficiency Measurement.

I. Introduction

Measuring the efficiency of organizations has become an important and appealing research area in recent years. Many public organizations depend on their income from public funds, thus making it essential in interest of accountability, to measure the efficiency of such institutions. These organizations are “non-profit”; thus there is an absence of output and input prices while producing multiple outputs from multiple inputs. This poses a challenge in measuring their efficiencies.

An assortment of methodological approaches has been employed in an effort to resolve the problem of efficiency measurement in this context. These include the deterministic frontier approach, the stochastic frontier approach and the mathematical programming approach. The latter approach differs from both statistical frontier approaches in that it is fundamentally non-parametric and from the stochastic frontier approach in that it is non-stochastic.

The mathematical programming approach differs from both statistical frontier approaches in that it is fundamentally non-parametric, and from the stochastic frontier approach in that is non-stochastic. There are a number of benefits implicit in the programming approach that makes it attractive on a theoretical level. Simulation studies have indicated that the piecewise linear production frontier formulated by DEA is generally more flexible in approximating the true production frontier than even the most flexible parametric function form [1].

DEA as originally proposed by A. Charnes, et al [2] is a non-parametric frontier estimation methodology based on linear programming for measuring relative efficiencies and performance of a collection of related comparable entities called Decision Making Units (DMUs) which transform multiple inputs into multiple outputs. The objective of a DEA study is to assess the efficiency of each DMU in relation to its peers. The result of a DEA study is a classification of all DMUs as either “efficient” or “inefficient”. Not only classifying the entities, but also determining the source of inefficiency (input) and the corresponding level or amount to enhance the performance. Unfortunately, the available DEA models and most of the previous DEA applications considering input variables of deterministic nature although some might have a random nature. Therefore in this study, we are looking forward to apply a Stochastic DEA since some of our input parameters have a random nature. Accordingly, we are proposing a modification to the existing stochastic DEA model which is chance constrained output oriented in order for the DEA model to be chance constrained input oriented.

The rest of the paper is organized as follows. The coming section discusses the methodology of the general DEA model and that for the stochastic DEA model. The third section includes the proposed stochastic chance constrained input oriented DEA model, the proposed solving algorithm, and a hypothetical illustrative example. The paper will end with the customary conclusions and implications for the future.

II. General Model of DEA

The basic DEA model for ‘n’ DMUs with ‘m’ inputs and ‘s’ outputs was first proposed by A. Charnes, et al [2]. The model determines the relative efficiency score for the different DMUs. The model depends on maximizing a production function estimated by DEA. This function is a deterministic frontier. For any inputs, the value of the DEA estimate defines the maximum output producible from inputs under all circumstances. On the other hand, for any outputs, the value of the DEA estimate defines the minimum input producing a given output under all circumstances. In this sense, it is comparable to the parametric frontier with one-sided deviations estimated using mathematical programming methods.

According to the assumptions relating the change in outputs as a result of the change in inputs, the DEA model can be classified as having either constant returns to scale (CRS) or variable returns to scale (VRS). Under CRS models the outputs are not affected by the size of the DMU, rather they change in direct proportion to the change in inputs assuming that the scale of operation does not influence efficiency; therefore, in the CRS models the output and input oriented measures of efficiency are equal. Under VRS models, changes in outputs are not necessarily proportional to the changes in the inputs; therefore In the VRS models the output and input oriented measures of efficiency scores are not equal for inefficient units [3]. In this paper we concerned about input oriented VRS model, the model is as follows:

$$\text{Min} \quad Z_p = \theta$$

$$\text{s.t.}$$

$$\sum_{i=1}^{n} \lambda_i x_i \le \theta x_p \qquad , \forall\, j = 1, \dots m$$

$$\sum_{i=1}^{n} \lambda_i y_i \ge y_p \qquad , \forall\, k = 1, \dots s \qquad (1)$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \ge 0 \qquad , i = 1, \dots n$$

III. The Chance Constrained output oriented Data Envelopment Analysis model

In a typical DEA model, the production function estimated by DEA is deterministic. Subhash Ray [4] modified the standard DEA model to measure relative efficiency in the presence of random variation in the all outputs produced from given inputs. For any input bundle, the value of the DEA estimate defines the maximum output producible from inputs under all circumstances. In the stochastic output oriented model, the inputs are assumed to be deterministic while all outputs are random, each output y_k is normally distributed with mean μ_p and variance σ_p^2 and the relation between the same stochastic output variable through different DMUs is independent, this means $cov(y_k, y_p) = 0$. The latter restriction regarding the output quantities in the DEA model translates into a random inequality that may at times be violated. Because an inequality involving a number of random variables can never be imposed with certainty, the strategy in CCP is to ensure that the probability that the inequality holds for a random sample of these variables does not fall below a certain level. Accordingly, the chance constrained output oriented model that measures the efficiency level of DMUp is as follow:

$$\text{Max}\, Z_p = \emptyset$$

$$s.t.$$

$$\sum_{i=1}^{n} \lambda_i \mu_i - \emptyset \mu_p \ge e \sqrt{\sum_{\substack{i=1 \\ i \ne p}}^{n} \lambda_i^2 \sigma_i^2 + \left(\lambda_p - \emptyset\right)^2 \sigma_p^2} \quad , \forall k = 1, \dots\dots s$$

$$\sum_{i=1}^{n} \lambda_i x_i \le x_p \qquad , \forall j = 1, \dots\dots m \qquad (2)$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \ge 0, (i = 1, 2, \dots., n)$$

where e is Significance level

IV. Developed Input Oriented Stochastic DEA Model

Since interested in evaluating the performance of public HEIs in order to assure the quality given that some of the input variables might have as stochastic nature, it was necessary to develop a stochastic input oriented model. Therefore, in this section we present our modification to the standard DEA model (Deterministic DEA) in order to measure technical efficiency in the presence of random variation in some of the inputs. Our developed Stochastic Input Oriented DEA model which is also based on the CCP method is provided below [5].

A. The Chance Constrained Input Oriented Data Envelopment Analysis Model

The restriction involving some of input quantities in the DEA model will be a random inequality that may at times be violated. Because an inequality involving a number of random variables can never be imposed with certainty, the strategy in CCP is to ensure that the probability that the inequality holds for a random sample of these variables does not fall below a certain level. The chance-constrained input oriented model for measuring the efficiency level of DMUp is as follow:

$$\text{Min}\, Z_p = \theta$$

$$s.t.$$

$$pr\left\{\sum_{i=1}^{n} \lambda_i x_i \le \theta x_p\right\} \ge \left(1 - \alpha_j\right) \qquad , \forall j = 1, \dots\dots J_S$$

$$\sum_{i=1}^{n} \lambda_i x_i \le \theta x_p \quad , \forall j = 1, \dots\dots J_D$$

$$\sum_{i=1}^{n} \lambda_i y_i \ge y_p \qquad , \forall k = 1, \dots\dots s \qquad (3)$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \ge 0, (i = 1, 2, \dots., n)$$

After that, we need to know nature of the relation between each DMU for each stochastic input variable, through statistical measure which is covariance. Covariance (cov) is a statistical measure of correlation of the fluctuations of two different quantities. The value of the covariance is interpreted as follows:

Positive covariance: implies that one variable is above (below) its mean value when the other variable is above (below) its mean value.

Negative covariance: implies that one variable is above (below) its mean value when the other variable is below (above) its mean value.

Zero covariance: if the two random variables are independent, the covariance will be zero. However, a covariance of zero does not necessarily mean that the variables are independent.

Assume that:

some of inputs are random variables and remaining inputs are deterministic variables;

each input $x_j, j \in J_S$ is normally distributed with mean μ_p and variance σ_p^2; and

the relation between the same stochastic input variable through different DMUs is dependent, this means $cov(x_i, x_p) \neq 0$.

Then, we can define a random variable *u*:

$$u = \sum_{i=1}^{n} \lambda_i x_i - \theta x_p \tag{4}$$

with mean:

$$E(u) = \sum_{i=1}^{n} \lambda_i \mu_i - \theta \mu_p \equiv \mu_u \tag{5}$$

and with variance:

$$var(u) = \sum_{\substack{i=1 \\ i \neq p}}^{n} \lambda_i^2 \sigma_i^2 + \left(\lambda_p - \theta\right)^2 \sigma_p^2 + 2cov\left(x_i, x_p\right) \equiv \sigma_u^2 \tag{6}$$

Since the $x_i's$ are normally distributed with mean μ_u and variance σ_u^2, therefore the variable *u* can be transformed into its equivalent standardized normal value *z*, as follows:

$$z = \frac{u - \mu_u}{\sigma_u} \tag{7}$$

Hence,

$$pr\left\{\sum_{i=1}^{m} \lambda_i x_i \leq \theta x_p\right\} = pr\{u \leq 0\} = pr\left\{z \leq \frac{-\mu_u}{\sigma_u}\right\} \tag{8}$$

Given the symmetric property of the normal distribution, then:

$$pr\left\{z \leq \frac{-\mu_u}{\sigma_u}\right\} = pr\left\{z \geq \frac{\mu_u}{\sigma_u}\right\} = 1 - \varphi\left(\frac{\mu_u}{\sigma_u}\right) \tag{9}$$

where $\varphi(\,)$ is the cumulative standard distribution function.

The random inequality restriction in the chance constrained DEA problem (CCDEAP) can be replaced by the equivalent restriction:

$$1 - \varphi\left(\frac{\mu_u}{\sigma_u}\right) \geq (1 - \alpha) \tag{10}$$

$$-\varphi\left(\frac{\mu_u}{\sigma_u}\right) \geq -\alpha \tag{11}$$

$$\varphi\left(\frac{\mu_u}{\sigma_u}\right) \leq \alpha \tag{12}$$

$$\varphi\left(\frac{\mu_u}{\sigma_u}\right) \leq \varphi(e) \tag{13}$$

$\varphi(e)$ is obtained from the table of standard normal distribution. Hence equation (13) can be written as,

$$\mu_u \leq e\sigma_u \tag{14}$$

Substitute equations (5) and (6) in equation (14). i.e.,

$$\sum_{i=1}^{n} \lambda_i \mu_i - \theta \mu_p \leq e \sqrt{\sum_{\substack{i=1 \\ i \neq p}}^{n} \lambda_i^2 \sigma_i^2 + \left(\lambda_p - \theta\right)^2 \sigma_p^2 + 2cov\left(x_i, x_p\right)} \tag{15}$$

Finally, from the above mathematical manipulation, the new presentation for the SIODEA model provided in (3) is as shown below:

$$\text{Min } Z_p = \theta$$

$$s.t.$$

$$\sum_{i=1}^{n} \lambda_i \mu_i - \theta \mu_p \leq \sum_{i=1}^{n} \lambda_i x_i \leq \theta x_p \quad ,\forall j = 1, \dots \dots . J_D$$

$$\sum_{i=1}^{n} \lambda_i y_i \geq y_p \quad , \quad \forall k = 1, \dots \dots . m \tag{16}$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \geq 0, (i = 1,2, \dots . , n)$$

B. The Algorithm of Developing SIODEA Model

From the previous section, we reached to the final form of the mathematical stochastic input oriented DEA model. Therefore the algorithm and its related flow chart (Figure 1) for the developing model will be as follows:

```
Step1: **Input**: $n$, $s$, $J_D$, $J_S$, $e$
Step2: **Set** $i$ =1
        while $i \le n$
            a. Input: deterministic inputs, parameters of
               stochastic inputs, deterministic outputs
            b. Set $\theta_i = 0$, $\lambda_i = 0$
            c. $i$ =$i$+1
       endwhile
Step3: **Set** $i$ =1
        while $i \le n$
            a. Formulate model for DMU$_i$
            b. Calculate $\theta_i$
            c. If $\theta_i = 1$ is true
                 Print "DMU$_i$: $\theta_i * 100\%$, Efficient"
               else
                 Print" DMU$_i$: $\theta_i * 100\%$, Inefficient"
              endif
            d. $i$ =$i$+1
       endwhile
```

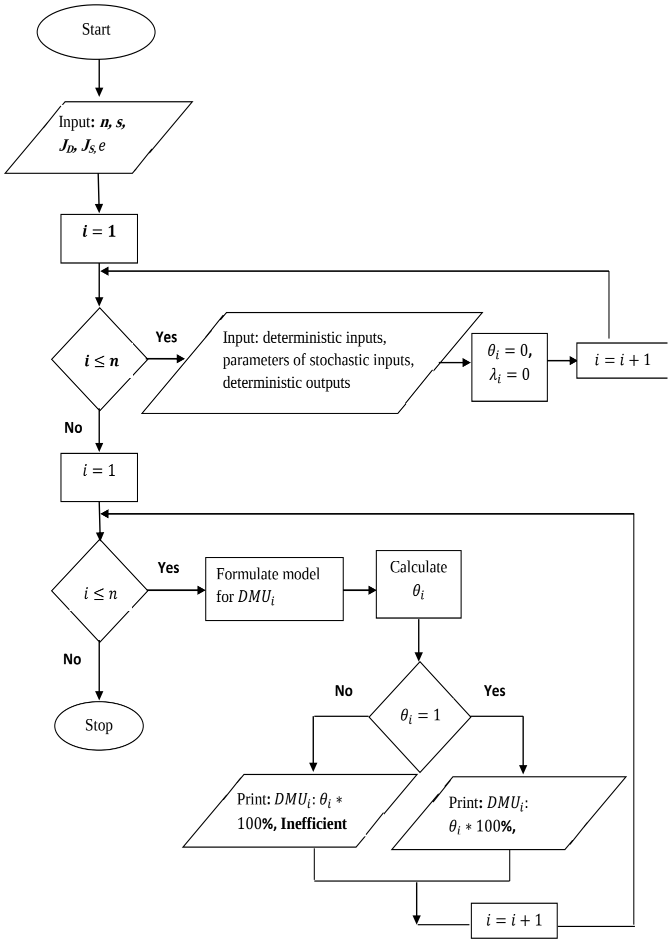

Start
Input: n, s, J_D, J_S, e
i = 1
i ≤ n
Yes
Input: deterministic inputs, parameters of stochastic inputs, deterministic outputs
θ_i = 0, λ_i = 0
i = i + 1
No
i = 1
i ≤ n
Yes
Formulate model for DMU_i
Calculate θ_i
No
Stop
No
θ_i = 1
Yes
Print: DMU_i: θ_i * 100%, Inefficient
Print: DMU_i: θ_i * 100%,
i = i + 1

Fig.1. The flow chart for SIODEA Model

C. Illustrative Example

Step1:

The following hypothetical example considers three universities with two input variables which are number of professors in each university (deterministic) and annual budget (stochastic), and three deterministic outputs which are the number of diplomas, bachelors, and masters granted in year by each university. Budget variable is normally distributed with different mean and variance for each university (DMU). The data for the deterministic variable is provided in Table I, while that for the parameters of stochastic variable is assumed in Table II.

Step2:

TABLE I. HYPOTHETICAL DATA FOR THE INPUT AND OUTPUTS FOR THE THREE UNIVERSITIES

University	Deterministic Inputs	Outputs		
	No. of Professors	***No. of Diploma***	***No. of Bachelors***	***No. of Masters***
A	5	9	4	16
B	8	5	7	10
C	7	4	9	13

TABLE II. HYPOTHETICAL DATA FOR THE STOCHASTIC INPUT (BUDGET) FOR THE THREE UNIVERSITIES

University	μ	σ^2	
A	14	1.4	$cov(x_A, x_B) = 0.9$
B	15	1.5	$cov(x_A, x_C) = 0.6$
C	12	1.2	$cov(x_B, x_C) = 0.7$

The aim of this problem is to determine the relative efficiency of the DMUs with respect to each other using the SIODEA model. Assume for the problem that the level of significance for the problem is 5%, and hence e will be 1.96.

Step3:

As has been explained earlier in model (16), a NLP formulation for each university has to be provided in order to measure the relative efficiency.

To evaluate the relative efficiency of ***university A***, we need to solve the following NLP problem:

$$\text{Min } Z_A = \theta$$

$$s.t.$$

$$14\lambda_A + 15\lambda_B + 12\lambda_C - 14\theta \le$$
$$1.96\left[\sqrt{1.5\lambda_B^2 + 1.4(\lambda_A - \theta)^2 + 2 * 0.9} + \sqrt{1.2\lambda_C^2 + 1.4(\lambda_A - \theta)^2 + 2 * 0.6}\right]$$

$$5\lambda_A + 8\lambda_B + 7\lambda_C \le 5\theta$$

$$9\lambda_A + 5\lambda_B + 4\lambda_C \ge 9$$

$$4\lambda_A + 7\lambda_B + 9\lambda_C \ge 4$$

(17)

$$16\lambda_A + 10\lambda_B + 13\lambda_C \ge 16$$

$$\lambda_A + \lambda_B + \lambda_C = 1$$

$$\lambda_A,\ \lambda_B,\ \lambda_C \ge 0$$

The first constraint is responsible for the representation of the stochastic input variable constraint, where the left hand side represents summation of means for each DMU minus the mean of university *A* and the right hand side represents the product of the level of significance and square root of summation of the terms variance of each DMU except university *A*, variance of university *A*, and twice covariance between university *A* and other DMUs. The second constraint is responsible for the representation of the deterministic input variable constraints, where the left hand side is the summation of input value for each DMU and the right hand side represents the product of input value of university *A* and efficiency of university *A*. The third, fourth and fifth constraints are responsible for representing the output variables constraints, where the left hand side of each constraint represents the summation of output value for each DMU and the right hand side represents the output value of university *A*. Finally, the sixth constraint ensures that the total weights for all DMUs equals to 1.

Similarly, the relative efficiency models for university *B* and *C* respectively are provided in (18) and (19) below.

University B:

$$\text{Min } Z_B = \theta$$

$$s.t.$$

$$14\lambda_A + 15\lambda_B + 12\lambda_C - 15\theta \le 1.96\left[\sqrt{1.4\lambda_A^2 + 1.5(\lambda_B - \theta)^2 + 2 * 0.9} + \sqrt{1.2\lambda_C^2 + 1.5(\lambda_B - \theta)^2 + 2 * 0.7}\right]$$

$$5\lambda_A + 8\lambda_B + 7\lambda_C \le 8\theta$$

$$9\lambda_A + 5\lambda_B + 4\lambda_C \ge 5$$

$$4\lambda_A + 7\lambda_B + 9\lambda_C \ge 7 \qquad (18)$$

$$16\lambda_A + 10\lambda_B + 13\lambda_C \ge 10$$

$$\lambda_A + \lambda_B + \lambda_C = 1$$

$$\lambda_A,\ \lambda_B,\ \lambda_C \ge 0$$

University C:

$$\text{Min } Z_C = \theta$$

$$s.t.$$

$$14\lambda_A + 15\lambda_B + 12\lambda_C - 12\theta \le 1.96\left[\sqrt{1.4\lambda_A^2 + 1.2(\lambda_C - \theta)^2 + 2 * 0.6} + \sqrt{1.5\lambda_B^2 + 1.2(\lambda_C - \theta)^2 + 2 * 0.7}\right]$$

$$5\lambda_A + 8\lambda_B + 7\lambda_C \le 7\theta$$

$$9\lambda_A + 5\lambda_B + 4\lambda_C \ge 4$$

$$4\lambda_A + 7\lambda_B + 9\lambda_C \ge 9 \qquad (19)$$

$$16\lambda_A + 10\lambda_B + 13\lambda_C \ge 13$$

$$\lambda_A + \lambda_B + \lambda_C = 1$$

$$\lambda_A,\ \lambda_B,\ \lambda_C \ge 0$$

We then used the GAMS programming language software to solve the above 3 models for each university independently. After running the software the relative efficiency of each university is:

DMU A: 100%,

DMU B: 82.7%, and

DMU C: 100%.

The results reveal that both universities *A* and *C* are efficient, while university *B* is inefficient. In other words, the outputs generated by university *B* are low given the high inputs for the university (8 professors and 15 units of budget). As noticed from Table I and II, university *B* has the highest inputs among the three universities and the lowest outputs.

V. Conclusion and Future Work

Data Envelopment Analysis is an excellent tool for the evaluation of performance. It has the advantage over alternative methods that it can be applied in a multiple inputs and outputs production context. A review of DEA applications revealed that most (if not all) has used the standard deterministic DEA model. They based their selection on the fact that the input and output variables are deterministic by nature, although some might be stochastic in nature. A new model, SIODEA, from the standard DEA model to handle random input variables was developed. The model considers that some of the inputs are stochastic following a normal distribution and the remaining inputs and all outputs are deterministic. Also, it is assumed that the covariance between the different DMUs within the stochastic input variable is not equal to zero. Through the example provided, the SIODEA showed promising results, and the model needs to be applied on actual studies.

As part of the future work, it is the intention of the authors to apply the developed SIODEA model to calculate and compare the efficiency of some public Egyptian universities. Further future work, is to develop a stochastic input oriented DEA model, where the limitations imposed on the stochastic variable are eliminated (i.e. the variables could follow any probability distribution). The model could then be expanded to include both stochastic input and output variables.

References

[1] Worthington, A., "An Empirical Survey of Frontier Efficiency Measurement Techniques in Education." Education Economics (2001) 9(3): 245-268.

[2] Charnes A., Cooper W., and Rhodes E., “Measuring the efficiency of efficiency of decision-making units”. European Journal of Operational Research. (1978) 2(6): 429-444.

[3] Banker R., Charnes A., and Cooper W., Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science. (1984) 30: 1078-1092.

[4] Subhash C. Ray, Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research. Cambridge University Press. (2004), pp. 307-325.

[5] El-Khodary, I., El-Demerdash, B., and Tharwat, A. “An Algorithm for Evaluating the Performance of Higher Education Organizations in Egypt Using a Stochastic DEA.” Proceedings of the 8th International Conference on Data Envelopment Analysis (DEA2010) - Performance Management and Measurement, American University of Beirut, Beirut, Lebanon. (2010).

Studies and a Method to Minimize and Control the Jitter in Optical Based Communication System

N. Suresh Kumar
GIT, GITAM Univetrsity,
Visakhapatnam, Andhrapradesh

Dr. D.V. R. K. Reddy
College Of Engineering, Andhra University,
Visakhapatnam

R. Sridevi
Asst Professor, Dept of ECE, Dr. Lankapalli Bullayya
College of Engineering for women,
Visakhapatnam

V. Sridevi
Sanketika Vidyaparishad Engineering College,
Visakhapatnam

***Abstract*—In the years, optical communication systems have been using significantly for attractive solutions to the increasing high data rate in telecommunication systems and various other applications. In the present days mostly, two types of communication schemes are using in data communication, namely asynchronous transmission and synchronous transmission depending on their timing and frame format. But both transmission systems are facing complications seriously with the involvement of jitter in data propagation. The jitter can degrade the performance of a transmission system by introducing bit errors and uncontrolled offsets or displacements in the digital signals. The jitter creates problems furiously at high data rate systems. The jitter need to be minimized in the communication system, otherwise it also degrades the performance of the interconnected systems with main circuit. This will happen due to improper synchronization or management of the clock scheme in the communication system. The improper organization of clock scheme propagates fault data and clock scheme to all other interconnected circuits. In the present work a new clock scheme is discussed to minimize the jitter in data propagation.**

Keywords—Optics; Jitter; pipeline; Clock; propagation delay; high speed data.

I. INTRODUCTION

Traditionally the jitter is measured in Unit Interval, where one Unit Interval corresponds to the phase deviation of one clock period. Controlling jitter is important because jitter can degrade the performance of a transmission system introducing bit errors and uncontrolled errors in the digital signals. Jitter causes bit errors by preventing the correct sampling of the digital signal by the clock recovery circuit in a regenerator or line terminal unit. The more the jitter grows, the smaller the valid bit interval at the end becomes, ultimately producing a higher bit error rate [9].

In optical fibre system the timing jitter generated by noise in the receiver and pulse distortion in the optical fibre. If the signal is sampled in the time between the signal crosses the threshold level, then the amount of distortion ΔT at the threshold level indicates the amount of jitter. Then the Timing jitter is given by,

Timing jitter (%) = $\Delta T/T_b X100\%$

Where, $T_b$is a bit interval
ΔT is the amount of distortion.

Traditionally, the rise time is defined as the time interval between the point where the rising edge of the signal reaches 10 percent of its final amplitude and the time it reaches 90 percent of its final amplitude. However, when measuring optical signals, these points are often obscured by noise and jitter effects. Thus, the more distinct values at the 20 percent and 80 percent threshold points are normally measured [12].

A similar approach is used to determine the fall time. Any nonlinearity of the channel transfer characteristics will create an asymmetry in the eye pattern. If a purely random data stream is passed through a pure linear system, all the eye openings will be identical and symmetrical.

II. METHOD OF TRANSMISSION

In parallel transmission, multiple channels are used to transmit several bits simultaneously, while a single channel is used in serial transmission. Data communication over short distances is generally using serial communication. It reduces the cost effect and complexity in interfacing by implementing single channel for communication. It also needed fewer devices in interfacing. The following parameters needed to consider for smooth transmitting of data over fibre link [12].

A. Line coding

The line coding is required to provide efficient timing recovery, synchronization as and suitable transmitted signal with less distortion. In the present paper optical fibre communications are used and hence large bandwidth is available. So, binary codes are mostly preferred for communication in the present paper. Basically, there are two types of two-level binary level codes that can be used for optical fibre transmission communication links. They are Return-Zero (RZ) and Non-Return-Zero (NRZ) format.

- **Return-Zero Coding (Transmitter)**

In Return to zero communication scheme the voltage levels return to zero after each transmission. In the RZ program, the microcontroller generates two types of clock pluses. One of the clock pulses is required for the shift register to serialize the parallel input from the USB module, another clock pulse is for the RZ module. The ratio of Shift register clock pulse to RZ clock pulse is 1:2.

- **Return Zero (Receiver)**

When the microcontroller detects the stop bit at the end of each byte, it will then generate 2 types of clock pulse. The first would be used to "push" the signal into the RZ circuit in which it will convert RZ signal into NRZ waveform. The second clock pulse will be used to read the signal from the serial to parallel shift register.

- **Non-Return-Zero Coding (Transmitter)**

The transmitter microprocessor consists of two programs, RZ and NRZ module. In the NRZ program, the microcontroller generates a type of clock pulse that is being used for shift register to serialize the parallel input from the USB. The serialize data is then feed to the data line for transmission. The transmitter signal is the same as the signal generated out from the shift register.

- **Non-Return-Zero Coding (Transmitter)**

When the microprocessor detects a start bit, it will start to tickle the shift register. The time duration for each clock pulse is approximately one micro second, which is approximately the same clock pulse, generated by the transmitter's clock. Thus one clock pulse is needed for the shift register to record in the data. After collecting all the 8 bits data, the microprocessor would finally generate a clock pulse to input the data from shift register to the computer via USB module.

B. Data communication scheme

Currently, two types of communication schemes are used in data communication. They are asynchronous transmission and synchronous transmission depending on their timing and frame format.

- **Asynchronous Transmission**

In asynchronous transmission, the transmitter and receiver clocks are free-running and are set to approximately the same speed. A start bit is transmitted at the beginning of each character, and at least one stop bit is sent at the end of character. The stop bit leaves the line or channel in the mark condition, which represents binary 1, and the start bit always switches the line to a space (binary 0). The timing remains accurate throughout the limited duration of the character as long as the clocks at the transmitter and receiver are reasonably close to the same speed. There is no set length of time between characters in asynchronous transmission. The receiver monitors the line until it receives a start bit. It counts bits, knowing character length being employed, and after the stop bit, it begins monitor the line again, waiting stop bit [13].

- **Synchronous Transmission**

In synchronous transmission, the transmitter and receiver are synchronized to the same clock frequency. As start and stop bits are not necessary, synchronous communication is more efficient than asynchronous and block of data is sent, that are much longer than a single character. Blocks begin with an identifying sequence of bits that allows proper framing and often identifies the content of the block. Synchronous transmission is more difficult and more expensive to implement than asynchronous transmission. It is used with higher transfer rates of communication: Ethernet etc. It is used in fast transfer rates (100kps to 100Mbps) [13].

The simplest solution for inter-domain data transfer is the two-flip-flop synchronizer [3]. The main problem with that synchronizer is its low throughput: typically, a complete transfer incurs waiting about one to two clock cycles at each end, and the next transfer cannot start before that handshake is complete. Although it is a very robust solution, it is sometimes misused or even abused in an attempt to reduce its latency [4][5].

In high-speed applications, logic channels are sometimes used as un-clocked information pipelines (e.g. wave pipelining), either to avoid distributing high-frequency clocks or to accommodate large delays (e.g. off-chip drivers) where synchronous pipelining is expensive or impractical [1][2]. In these circuits the quality is determined by clock skew and jitter [9]. In the conventional pipeline system, it is facing problems due to improper synchronization of clock pulses. This is a universal problem in all the digital systems mostly called clock skew. The system clocking must be such that the output data is clocked after the latest data has arrived at the outputs and before the earliest data from the next clock cycle arrives at the outputs [8].

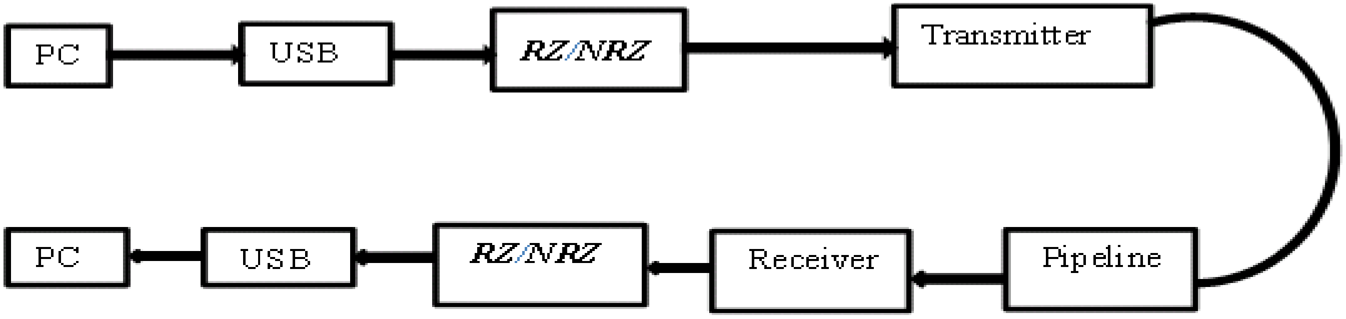


Fig.1. Block Diagram Optical Communication System with new method including pipeline.

In this crucial period it is difficult and highly impossible to get exact input data match with the output in conventional circuits. The problem can be solved with new clock scheme [8]. In the present paper same clock scheme with more intelligent system is adopted. In this new method, after arriving of first valid data at interrupt controller the interrupt controller interrupts microcontroller. In response to this interrupt the microcontroller send a clock signal to the next stage register.

Similarly after receiving a valid signal from second register the interrupt controller again interrupts the microcontroller. The microcontroller in the same way activates the next stage in a different path [10][11].

III. Motivation

In any Interfacing system the data must be fed to processor at exact clock pulse. If the handshaking does not exist between two processor and transmitter, there may be a chance of data loss. It leads degradation in accuracy [11]. For example in figure 1 if the transmitter speed is high than the receiver system, the receiver may loss some data. This is due to speed miss-match between transmitter and Receiver. There are some methods effectively acting to minimize these data losses using new pipeline techniques [11][10]. The new clock schemes and constraints in the new pipeline systems are described in the past methods [10]. In the present paper the new clock scheme is interfaced between transmitter and receiver to minimize the jitter.

IV. Experimental setup

An USB is used to interface the transmitter for fast and simple interface. The RZ and NRZ circuits are used in communication to minimize the power consumption and length of sampling time. Figure 2 is showing Return Zero (RZ) and Non-Return Zero module used for transceiver communication. A detailed description of outputs and operations are discussed in the next section. The inputs of various ranges are produced by function generator, to feed the circuit as shown in figure 2. The inputs are feed to RZ /NRZ circuits through USB. An optical fiber is connected for data transmission. The optical fiber is used as channel to carry the data pulses to satisfy the operations through transceiver. Optical fiber is selected in the present work because of its vast advantage in tele-communication system. Optical fiber supports fast data transmission rates. But it generates jitter due to noise at receiver side and due to the distortion in the fiber. A two stage pipeline is interfaced between transmitter and receiver to minimize the jitter and data losses.

Advantages of Optical Fiber: Some of the advantages are discussed in the present section.

a) In the long distance communication the network connection is more flexible and transmission errors are almost zero.

b) Fiber cable support higher data rates. That is why, in the present work a pipeline technique is used to synchronize the transmitter with receiver.

c) Fiber cables have long life time when compare with copper wires. They are more reliable than any other channel.

d) Fiber cables are resistible to cross talk and Electro Magnetic Interferences.

e) They are easier to test and interface.

f) Installation costs of fiber cables are cheaper than other channel installations.

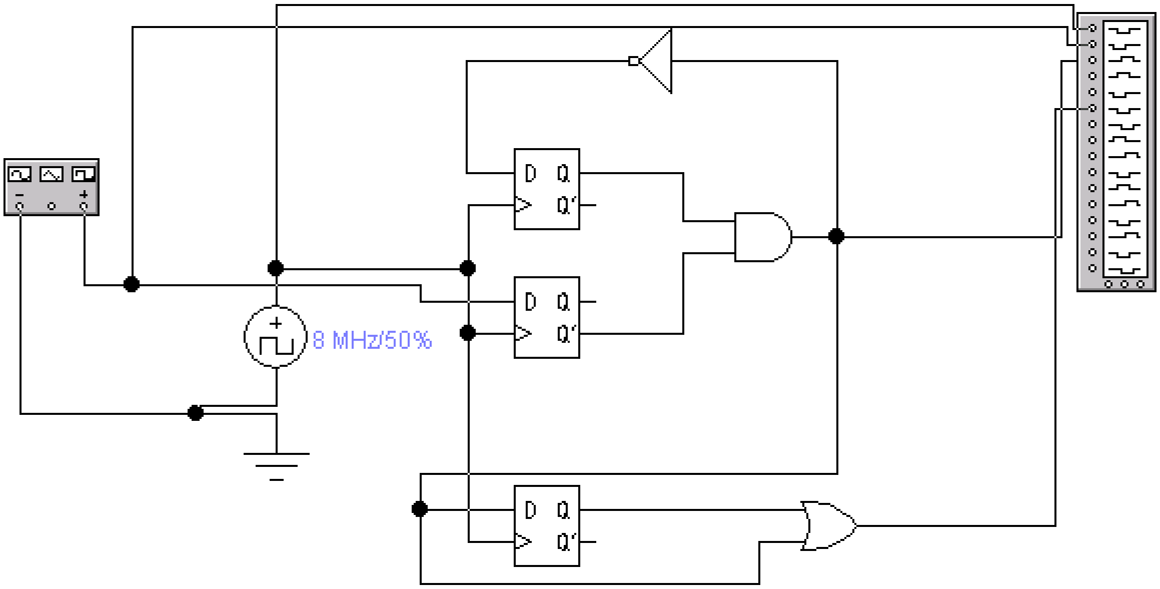
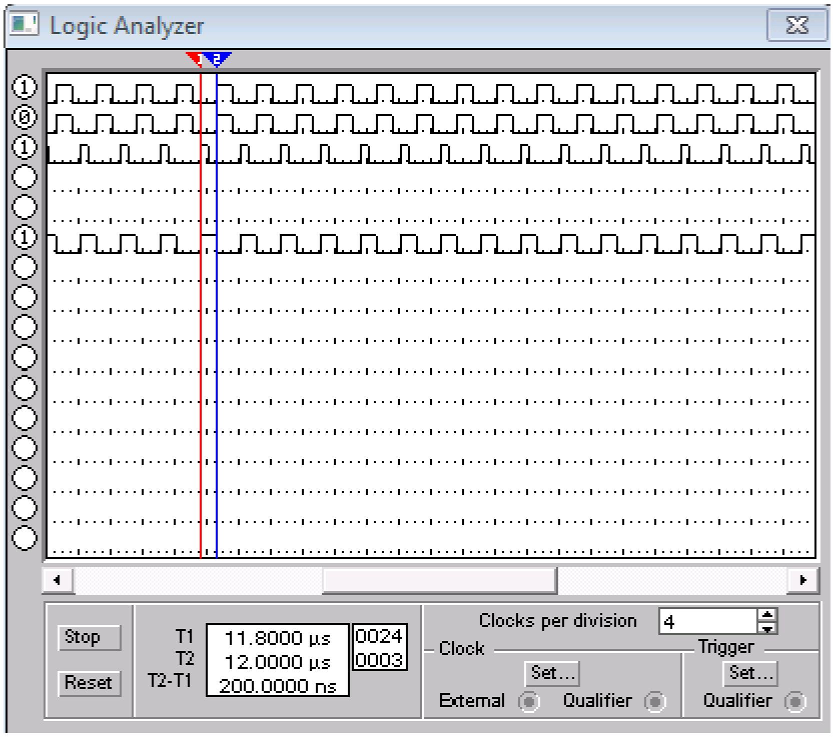


Fig.2. Analysis Circuit of Return Zero Transmitter and Receiver Module

V. Results

In figure 3 and figure 4 the first wave showing the clock pulse applied to the circuit. The second pulse is generated from function generator given as input data to RZ and NRZ circuit. The red line and blue vertical lines are representing the width of the pulses. The outputs are generated and analysed in Electronic work bench and Proteus. The circuit is analysed for various ranges of input data rates [11] and taken screen shots as shown in figure 3 to figure 7. The Digital circuit is studied without pipeline and then after compared the operation by integrating with pipeline in the circuit.

At higher data rates, due to noise and jitter some distortions are observed as shown in figure 6. By interfacing the new pipeline module some distortions are minimized, which is shown in figure 7. In figure 6 and figure 7, the yellow pulse represents the first stage clock pulse and blue line represents the second stage clock pulse. The pink pulse represents the data output of the optical fiber and green colour represents the data output at the receiver. Figure 6 is the simulated output of the circuit without pipeline and figure 7 is the simulated output of the circuit with pipeline.

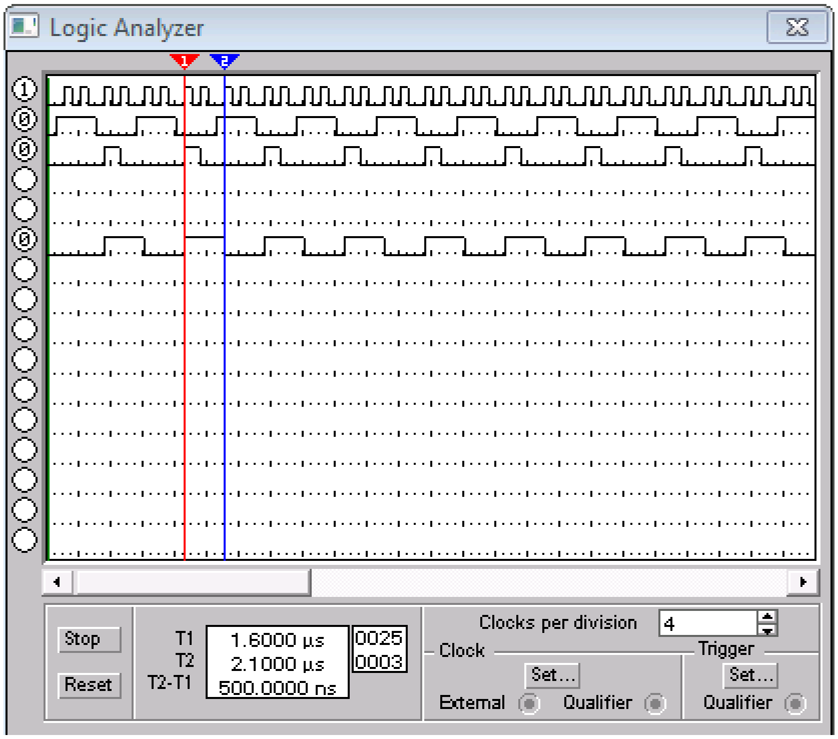


T1 11.8000 µs
T2 12.0000 µs
T2-T1 200.0000 ns
0024
0003

Clocks per division 4

Clock: Set...; External ○ Qualifier ○
Trigger: Set...; Qualifier ○

Fig.3. Logic Analyser of Return Zero Transmitter and Receiver Module with 2MHz input

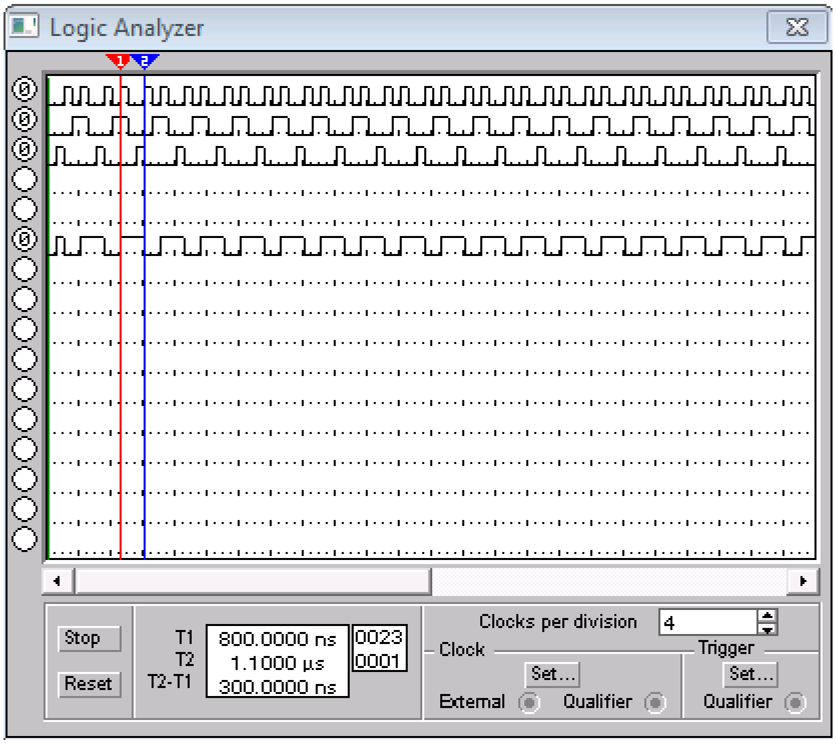


T1 1.6000 µs
T2 2.1000 µs
T2-T1 500.0000 ns
0025
0003

Clocks per division 4

Clock: Set...; External ○ Qualifier ○
Trigger: Set...; Qualifier ○

Fig.4. Logic Analyser of Return Zero Transmitter and Receiver Module with 1MHz input



Logic Analyzer

T1 800.0000 ns 0023
T2 1.1000 µs 0001
T2-T1 300.0000 ns

Clocks per division 4

Fig.5. Logic Analyser of Return Zero Transmitter and Receiver Module with 8MHz input

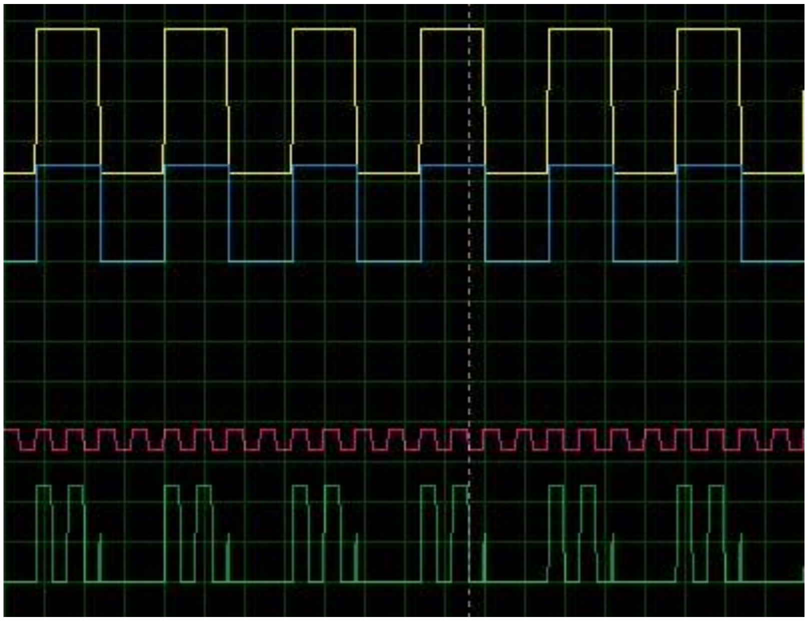


Fig.6. Logic analyser of Optical communication system without pipeline

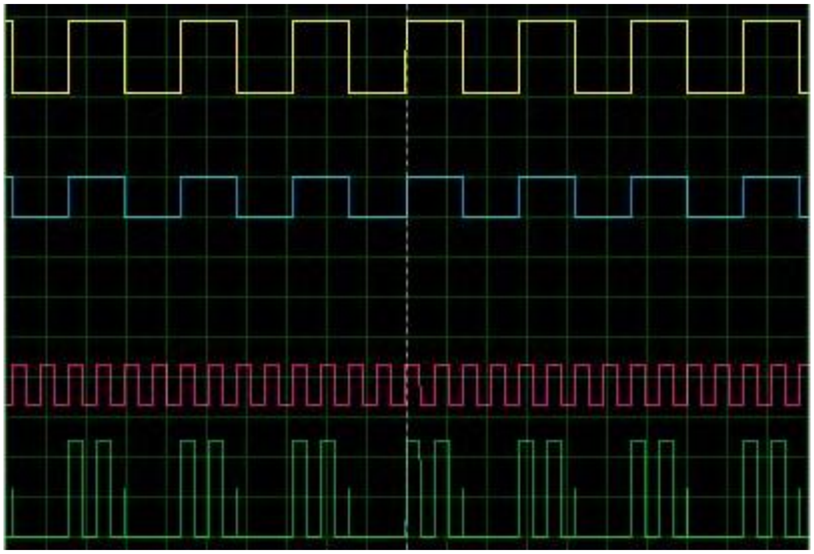


Fig.7. Logic Analyser of optical communication system with new pipeline system

At higher frequency rates the RZ receiver is unable to produce the desired output. At high frequencies the pulse width is more when compared with low frequency rate receiver module. This deviation produces distortion at receiver. In the present work a new pipeline technique is discussed to minimize jitter at receiver side. A new pipeline technique [1] is interfaced between optical system and receiver RZ module.

VI. Conclusion

Jitter can degrade the performance of the fibre-optic system by causing bit errors. The faster and more complex the system becomes, management of jitter is increasingly critical. This is because; at higher data rates bits are placed more closely together. In the present work the jitter minimization techniques are discussed to improve the system performance at wide range of data rates. The system is clocked such that a pipeline stage is operating on more than one pulse simultaneously. The present system at any given time, multiple pulses can be present in a stage.

References

[1] Dally WJ, Poulton JW (1998) Digital systems engineering. Cambridge University Press

[2] Dike C, Burton E (1999) Miller and noise effects in a synchronizing Flip-flop. IEEE J. Solid-State Circuits 34(6):849–855

[3] Kinniment DJ, Bystrov A, Yakovlev A (2002) Synchronization circuit performance. IEEE J. Solid-State Circuits 37:202–209

[4] Ginosar R (2003) Fourteen ways to fool your synchronizer. Proc. 9th IEEE Int. Symp. on Asynchronous Circuits and Systems (ASYNC03)

[5] Uri Frank et al., "A predictive synchronizer for periodic clock domains", Form Method Syst Des (2006) 28: 171–186, DOI 10.1007/s10703-006-7843-9, Springer publications.

[6] L W.C Ouen, "Maximum-rate pipeline systems". 1969 MIPS Proceedings.

[7] Derek Wong, Giovannc De Micheli and Michael Flynn. "Inserting activeings of Spring Joint Computer. pp. 5816. delay elements to achieve wave pipelirung," ICCAD 1989. w. 270-273. 1989.

[8] N. Suresh Kumar et al., "Method to Minimize Data Losses in Multi Stage Flip Flop", GJRE (F), Volume 11 Issue 6 Version 1.0 November 2011

[9] Pemg-Shyong Lin et al., "Jitter Due to Signal History in Digital Logic Circuits and its Control Strategies", doi: 0-7803-1254-6/93$03.0Q0 1993 IEEE

[10] N. Suresh Kumar et al., "Effect of Interrupt Logic on Delay Balancing Circuit", IJCA, *Volume 27– No.4, August 2011.*

[11] N. Suresh Kumar et al., "A New Method to Enhance Performance of Digital Frequency Measurement and Minimize the Clock Skew", IEEE Sensors Journal, VOL. 11, NO. 10, October 2011, pp 2421-2425..

[12] John M. Senior, "Optical Fibre Communications", Prentice- Hill Inc., 1999.

[13] William Stallings," Data and Computer communications". Macmillan, 1991.

Biography

N. Suresh Kumar received his B.E. degree from Berhampur University, India, in 2001 and M.Tech. degree from Allahabad Deemed University, India, in 2005. Currently, he is working as a research scholar in Andhra University, Visakhapatnam, India. He is currently working in GITAM University, Visakhapatnam, India. During his 10 years of experience he occupied different positions in academic and administration. His research interest is sensor measurement. He is also interested in developing new teaching methodologies in the class room. Some of his education related papers are published in International Journals and have been selected in some of the National and International Conferences. Some of his research papers are also published in various International journals.

D. V. Rama KotiReddy received the Ph.D. degree from Instrument Technology Engineering, Andhra University, Visakhapatnam, India, in 1996. He is coordinator for MEMS Design Laboratory, Andhra University. Throughout his 10+ years of professional experience, he shared his wisdom with fellow engineers and scientists through his valuable research papers published in international and National journals and Conference proceedings. He extended his guidance in the research fields of sensor networking, MEMS technologies, energy studies, and VLF communication.

He is influenced and actively doing current research work on MEMS technologies. His expertise is in providing engineering solutions and designs to many industrial companies. He extended his experience with UGC MHRD and INUP programme from IISc, Bangalore, to establish scientific Labs in the Instrument Technology Department, Andhra University. His services also extended as Associate Dean, Industrial Consultancy Cell, College of Engineering, Andhra University. He is Member Board of Studies, AU College of Engineering. He also contributes as resource person in many national and International conferences. Dr. Rama KotiReddy was a recipient of a Senior Research Fellowship from UGC.

R. Sridevi has received her M. Tech degree from Andhra University, Visakhapatnam. Presently she is working as assistant professor in ECE dept of Dr. Lankapalli Bullayya College of Engineering for women, Visakhapatnam. Some of her research papers are also published in various International journals. Her research of interests are MIMO, Radar and microwave engineering, Sensor networking. She has served as technical advisor to, many engineering college Laboratories.

V. Sridevi has received her Diploma in Electronics and Communication engineering in 2003. She has completed her B. Tech Degree from Andhra University in 2011. Presently she is doing her M. Tech. During her Professional experience she has published couple of papers in international journals. Her research of interests are MIMO, Radar engineering, Sensor networking. She has served as technical advisor to, many engineering college Laboratories.

Novel Steganography System using Lucas Sequence

Fahd Alharbi
Faculty of Engineering
King Abdulaziz University
Rabigh, KSA

***Abstract*—Steganography is the process of embedding data into a media form such as image, voice, and video. The major methods used for data hiding are the frequency domain and the spatial domain. In the frequency domain, the secret data bits are inserted into the coefficients of the image pixel's frequency representation such as Discrete Cosine Transform (DCT) , Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) . On the other hand, in the spatial domain method, the secret data bits are inserted directly into the images' pixels value decomposition. The Lest Significant Bit (LSB) is consider as the most widely spatial domain method used for data hiding. LSB embeds the secret message's bits into the least significant bit plane (Binary decomposition) of the image in a sequentially manner . The LSB is simple, but it poses some critical issues. The secret message is easily detected and attacked duo to the sequential embedding process. Moreover, embedding using a higher bit plane would degrade the image quality. In this paper, we are proposing a novel data hiding method based on Lucas number system. We use Lucas number system to decompose the images' pixels values to allow using higher bit plane for embedding without degrading the image's quality. The experimental results show that the proposed method achieves better Peak Signal to Noise Ratio (PSNR) than the LSB method for both gray scale and color images. Moreover, the security of the hidden data is enhanced by using Pseudo Random Number Generators (PRNG) for selecting the secret data bits to be embedded and the image's pixels used for embedding.**

Keywords—Steganography; LSB; Lucas; PSNR; PRNG

I. INTRODUCTION

The goal of the data hiding system is to communicate a secret message in a way that would not be noticeable by an intruder [1-3]. There are two techniques are regularly used for data hiding, the frequency domain [4-5] and the spatial domain [6-10]. In the frequency domain data hiding techniques, the secret data bits are inserted into the coefficients of the image pixel's frequency representation.

Among the frequency image pixel's frequency representation are Discrete Cosine Transform (DCT) [11], Discrete Fourier Transform (DFT) [12] and Discrete Wavelet Transform (DWT) [13]. On the other hand, in the spatial domain techniques, the secret data bits are inserted directly into the images' pixels value decomposition. In this section, we are discussing the Lest Significant Bit (LSB) which is consider as the most widely spatial domain method used for data hiding.

The least significant bit (LSB) data hiding technique [6-10] is the process of inserting the secret message's data bits into the least significant bits of the image (Binary Form) in a sequential manner [14-17].

For illustration, let I be the original gray scale image where each pixel is represented using 8-bit format, thus each pixel's value is varies from 0 up to 255 as illustrated at Table 1.

TABLE I. ORIGINAL IMAGE PIXELS (BINARY)

Pixels before embedding	
1st	01001100
2nd	01001101
3rd	01001110
4th	01001111
5th	01010000
6th	01001011
7th	01010001
8th	01010001

The secret message is simply the letter Z with its binary representation as 01011010. The LSB embedding process is shown at Table 2, where 8 pixels are used to hide the letter Z. The secret message's bits are inserted into the least significant bit of the image's pixels in a sequential manner.

The LSB data hiding technique is simple and the effect on the image quality is limited and hardly noticed by the human eye duo to the small value of the bit (least significant bit) used for embedding. On the other hand, the LSB is easy to be detected and attacked by simply extracting or changing the least significant bits of each pixel.

TABLE II. LSB EMBEDDING

Secret bits (Z)	Pixels after embedding
0	0100110**0**
1	0100110**1**
0	0100111**0**
1	0100111**1**
1	0101000**1**
0	0100101**0**
1	0101000**1**
0	0101000**0**

On the other hand, using higher bits for embedding the secret message would enhance the security and at the same time degrade the image quality. For example, as illustrated at Table 3, using the fifth bit for hiding the secret message would degrade the image quality duo to the fact that the value of the fifth bit is 16 and the impact would be clear.

TABLE III. LSB EMBEDDING (USING FIFTH BIT)

Secret bits (Z)	Pixels after embedding	difference
0	010**0**1100	0
1	010**1**1101	16
0	010**0**1110	0
1	010**1**1111	16
1	010**1**0000	0
0	010**0**1011	0
1	010**1**0001	0
0	010**0**0001	-16

In this paper, we propose using Lucas number system for image pixel's value decomposition to allow using higher bit plane without degrading the image quality. Also, we are enhancing the security of the data hiding system by using Pseudo Random Number Generator to select the next pixel used for embedding. The rest of the paper is organized as follows: Section II discusses the Lucas based hiding system; Section III describes enhancing the data hiding system's security by using Pseudo Random Number Generators to select the next pixel for embedding; Section IV presents experimental results; we finally conclude in Section V.

II. LUCAS BASED HIDING SYSTEM

Now, we are proposing using the Lucas numbers [18-19] for pixels values decomposition. The Lucas sequence generated using the following formula

$$L_n = L_{n-1} + L_{n-2} \quad , n > 2 \tag{1}$$

Where, $L_1 = 2$ and $L_2 = 1$

The image's pixels values would be represented as the sum of the non-consecutive Lucas numbers [20-21]. To represent the range of 0 to 255, we need 12-bit of Lucas digits.

L_{12}	L_{11}	L_{10}	L_9	L_8	L_7	L_6	L_5	L_4	L_3	L_2	L_1
199	123	76	47	29	18	11	7	4	3	1	2

Now, we represent the pixel value of 26 using Binary and Lucas representation as follows

The binary representation is 00011010
The Lucas representation is 000001010010

Let consider hiding the bit value of 0 using the fifth bit in each decomposition of the pixel value of 26. The result is as following

The binary representation is 000**0**1010
The Lucas representation is 0000010**0**0010

The pixel value after LSB embedding using the fifth bit is 10, while the pixel value after embedding using Lucas decomposition is 19. It is clear that using the Lucas decomposition would result in higher quality data embedding duo to the fact that the Lucas bits are less significant than those of the Binary decomposition.

Here, we reconsider the example of hiding the letter Z in eight pixels (Table 1). The eight pixels values in Lucas system are shown at Table 4.

TABLE IV. ORIGINAL IMAGE PIXELS (LUCAS)

Pixels before embedding	
1st	001000000000
2nd	001000000010
3rd	001000000001
4th	001000000100
5th	001000001000
6th	000101010100
7th	001000001010
8th	001000001010

Table 5 shows the result of embedding the letter Z into eight pixels using Lucas decomposition, where the image quality after embedding is better than that achieved using LSB technique duo to the fact that the digits in Lucas system are less significant (7) than those in binary system (16).

TABLE V. LUCAS EMBEDDING (USING FIFTH BIT)

Secret bits (Z)	Pixels after embedding	difference
0	0010000**0**0000	0
1	0010000**1**0010	7
0	0010000**0**0001	0
1	0010000**1**0100	7
1	0010000**1**0000	3
0	0001010**0**0100	-7
1	0010000**1**0010	3
0	0010000**0**1010	0

III. PSEUDO RANDOM NUMBER GENERATOR

The LSB technique is simple but not secure. The intruder can easily recover the hidden message by extracting the least significant bits. On the other hand, using higher bits for embedding would degrade the image quality. In this section, we enhance the data hiding system's security by using Pseudo Random Number Generators to select the next pixel for embedding. The Pseudo Random Sequence is generated using Non-Linear forward feedback shift Register (NLFFSR) [22-23]. To start generating the Pseudo Random Sequence, the registers simply loaded with any initial value except zero and with each clock (step) a new Random Number is generated. The feedback function of the Pseudo Random Sequence Generator is designed based on the characteristic polynomial of the Generator. The characteristic polynomial is in the form of

$$P(x) = x^n + a_{n-1}x^{n-1} .... + a_1x + a_0 \tag{2}$$

Where, n represents the number of registers and the length of the generated sequence is

$$N = 2^n - 1 \tag{3}$$

Let consider a gray scale image $I_{R,C}$, where C is the number of coulombs and R is the number of pixels (rows) in each coulomb.

To increase the security of the data hiding system, we use three Random Sequence Generators to select the next pixel and bit plane used for embedding $I_{i,j}^{b}$. The first Generator selects a coulomb j in the image, the second Generator selects a pixel (row) i at the selected coulomb and the third Generator selects the bit plane ($b = 1,2.....8$) used for embedding. On the other hand, the color image pixel is represented by three colors Red, Green, and Blue.

Each color is represented by 8-bit, thus each color's value varies from 0 up to 255. In this case, we use four Random Sequence Generators to select the next pixel, color and bit plane used for embedding $I_{i,j,k}^{b}$, where k ($k = 1,2,3$) is the selected color. Moreover, we may use two Random Sequence Generators to select the byte y of the secret data and the bit x to be embedded, S_{y}^{x} . The secure data hiding system may uses up to six Random Sequence Generators to perform the data embedding process as the following

$$I_{i,j,k}^{b} = S_{y}^{x} \tag{4}$$

Now, we enhance the embedding of the letter Z into the eight pixels by using Random Sequence Generators. Let assume that the coulomb j has been selected and it is contain the pixels values in the sequence shown at Table 1 and the sequence of the rows i (pixel) are in the following random sequence (6,8,3,5,2,4,7,1). The embedding process is shown at Table 6, where both the image quality and the data security are enhanced.

TABLE VI. LUCAS EMBEDDING + PRNG (USING FIFTH BIT)

Secret bits (Z)	Pixels after embedding	difference
0	0001010**0**0100	-7
1	0010000**1**0010	3
0	0010000**0**0001	0
1	0010000**1**0000	3
1	0010000**1**0010	7
0	0010000**0**0100	0
1	0010000**1**0010	3
0	0010000**0**0000	0

IV. EXPERIMENTAL RESULTS

In this section, we are evaluating the performance of the LSB data hiding technique and the proposed lucas based data hiding technique. The evaluation is performed by using a gray scale 512 × 512 pixels image and a color 512 × 512 × 3 image for data hiding. The quality of the embedding techniques are evaluated by the Peak Signal to Noise Ratio (PSNR), where it is defined as

$$PSNR = 10\log_{10}\left(\frac{I_{MAX}^{2}}{MSE}\right) \tag{5}$$

Where, I_{MAX}^{2} is equal to 255 as the maximum possible value of a pixel in the gray scale images or represents the maximum possible value of a color in the color images. The Mean Square Error (MSE) for the gray scale images is computed as follows

$$MSE = \frac{1}{R \times C}\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\left|I_{i,j} - I'_{i,j}\right|\right)^{2} \tag{6}$$

Where, $I_{i,j}$ is the original image's pixel value and $I'_{i,j}$ is the image's pixel value after embedding.

Also, The Mean Square Error (MSE) for the color images is computed as follows

$$MSE = \frac{1}{R \times C \times L}\sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{L}\left(\left|I_{i,j,k} - I'_{i,j,k}\right|\right)^{2} \tag{7}$$

Where, L is the number of colors in the color image pixel.

A. Performance Evaluation using Gray Scale Image

In this experiment, the performance of the data hiding techniques is evaluated using different bit planes for embedding the secret message's bits. The hiding capacity of the 512 × 512 pixels original image is 32768 data bytes, where each pixel is used for hiding a single data bit. We vary the bit plane used for embedding from the first bit up to the eighth bit for each data hiding technique. In each case we hide 32768 data bytes and evaluate the performance by computing the PSNR.

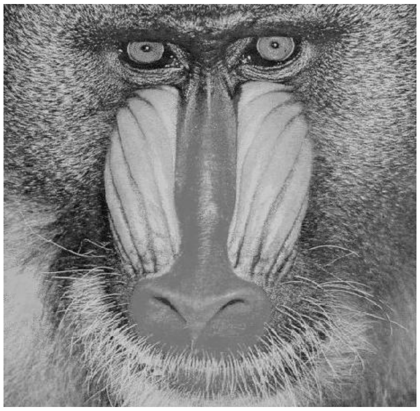
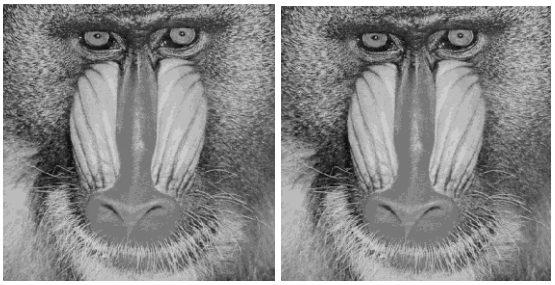


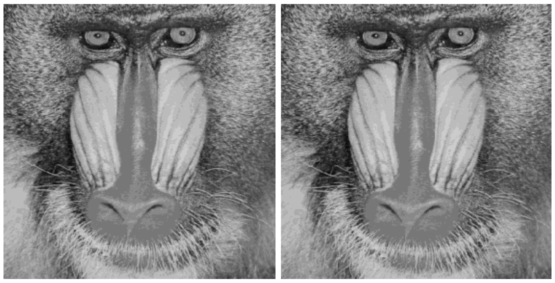
Fig.1. ORIGINAL GRAY SCALE IMAGE

The performance of the Embedding Techniques is shown at Figures (2-9). The quality of the covered image using the LSB hiding technique is degrade using higher bit plane for embedding. On the other hand, Lucas hiding technique maintains a better image quality for all cases. Moreover, Table 7 shows that Lucas hiding technique outperforms the LSB in achieving better Signal to Noise Ratio (PSNR) duo to the fact that digits in Lucas number system are less significant than those in binary system. Conversely, the first bit in the binary system is less significant than that of the Lucas number system , thus the LSB achieves better PSNR than the Lucas hiding technique with using the first bit for imbedding which is consider the least security bit plane for data hiding.



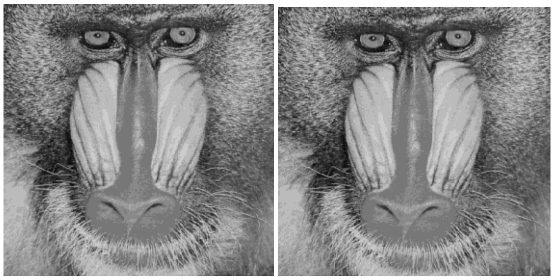
LSB LUCAS

Fig.2. EMBEDDING TECHNIQUES PERFORMANCE (1ST BIT)



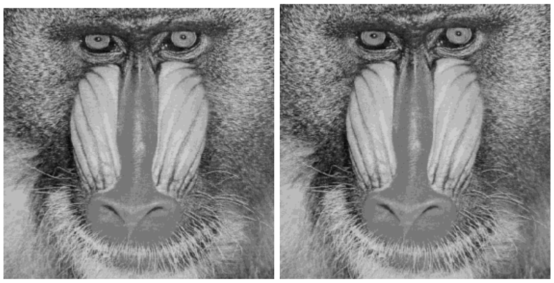
LSB LUCAS

Fig.3. EMBEDDING TECHNIQUES PERFORMANCE (2ND BIT)



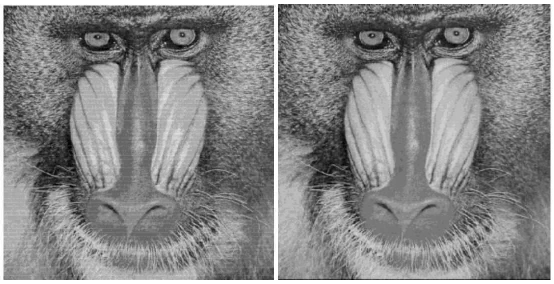
LSB LUCAS

Fig.4. EMBEDDING TECHNIQUES PERFORMANCE (3RD BIT)



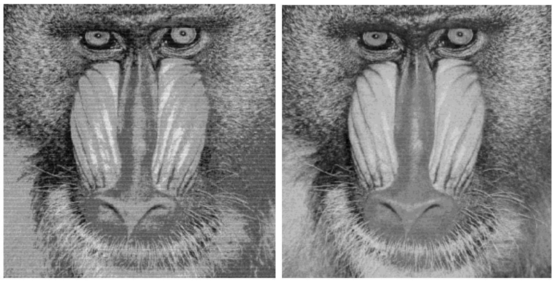
LSB LUCAS

Fig.5. EMBEDDING TECHNIQUES PERFORMANCE (4TH BIT)



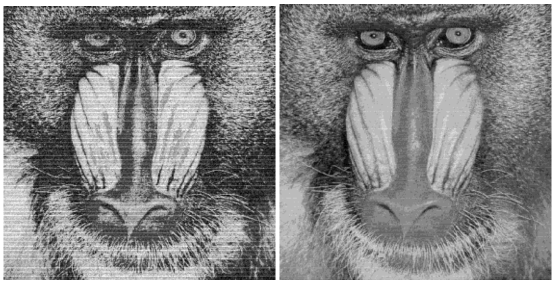
LSB LUCAS

Fig.6. EMBEDDING TECHNIQUES PERFORMANCE (5TH BIT)



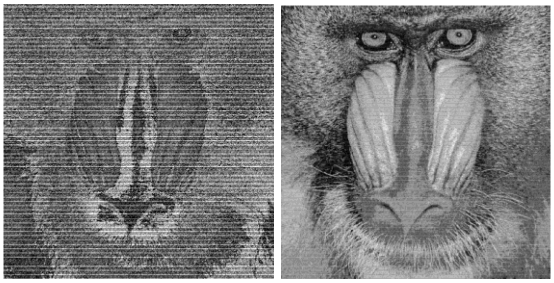
LSB LUCAS

Fig.7. EMBEDDING TECHNIQUES PERFORMANCE (6TH BIT)



LSB LUCAS

Fig.8. EMBEDDING TECHNIQUES PERFORMANCE (7TH BIT)



LSB LUCAS

Fig.9. EMBEDDING TECHNIQUES PERFORMANCE (8TH BIT)

TABLE VII. EMBEDDING TECHNIQUES PERFORMANCE (PSNR)

Bit plane	**Peak Signal to Noise Ratio (PSNR)**	
	LSB	**Lucas**
1st	77.0399	54.6944
2nd	50.7404	64.1260
3rd	41.7281	52.1163
4th	35.4584	43.6703
5th	27.3609	38.8316
6th	20.522	33.4006
7th	14.9971	27.7555
8th	8.9157	22.72

B. Performance Evaluation using Color Image and PRNG

In this experiment, the performance of the data hiding techniques are evaluated using a color 512 × 512 × 3 image (Figure 10) for data hiding. The hiding capacity of the original image is 98304 data bytes, where each pixel is used for hiding three data bits, one bit in each color. The embedding process

(Eq. 4) is performed using five Pseudo Random Sequence Generators. The first Generator selects a secret data byte y from the 98304 data bytes and the second Generator picks the bit x to be embedded in the following sequence ($x = 5,7,8,1,2,4,3,6$). The third and the fourth Generators select the position (coulomb j & raw i) of the pixel to be used for embedding as shown at Figures (11-12). The fifth Generator selects the color k in the following sequence ($k = 2,1,3$). Finally, we vary the bit plane b used for embedding from the first bit up to the eighth bit and in each case we hide 98304 data bytes. For performance evaluation, we compute the PSNR for each case. The performance of the LSB Embedding Technique is shown at Figures (13-20). The higher the bit plane used for embedding the higher the impact on the covered image quality. On the other hand, the data hiding system using Lucas numbers and PRNGs maintained a better image quality and enhanced the data security as illustrated at Table 8.

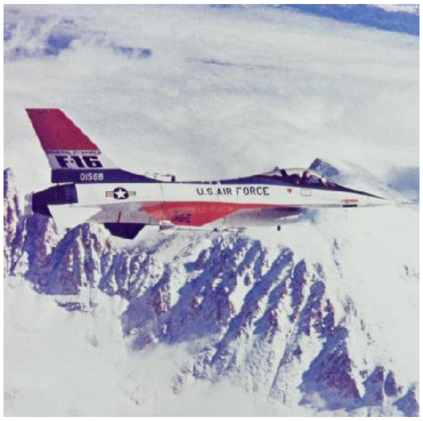


Fig.10. ORIGINAL COLOR IMAGE

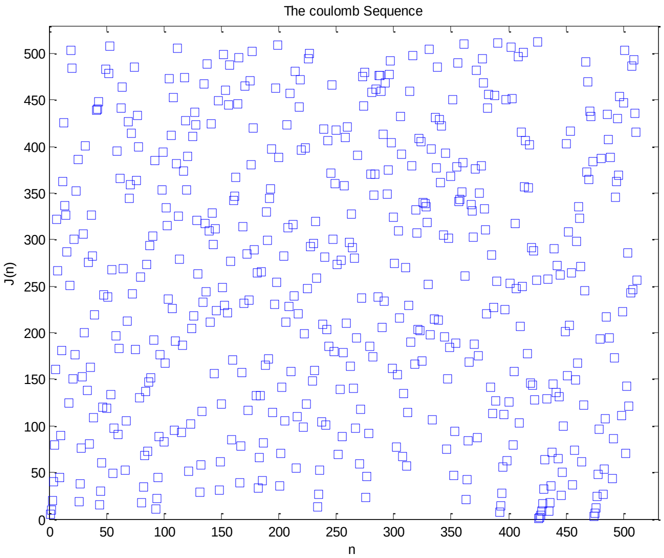


Fig.11. THE THIRD GENERATOR

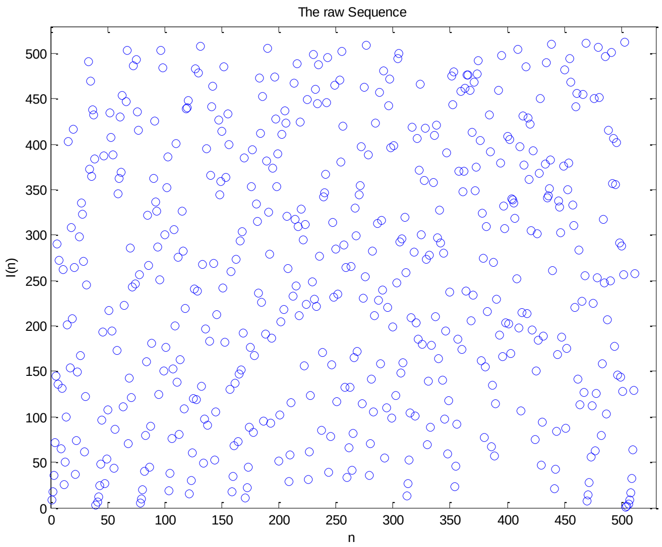
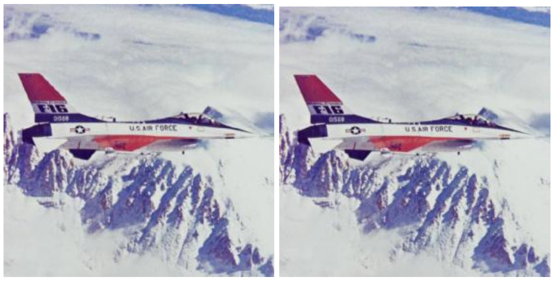
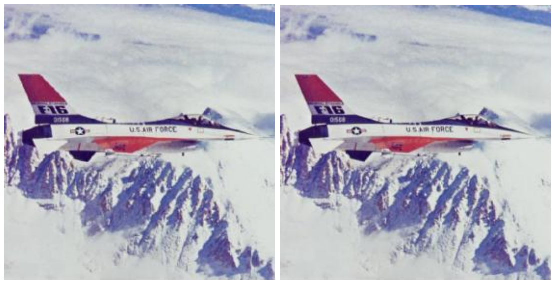


Fig.12. THE FOURTH GENERATOR



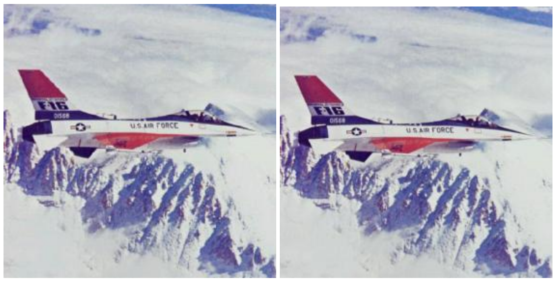
LSB LUCAS + PRNG

Fig.13. EMBEDDING TECHNIQUES PERFORMANCE (1ST BIT)



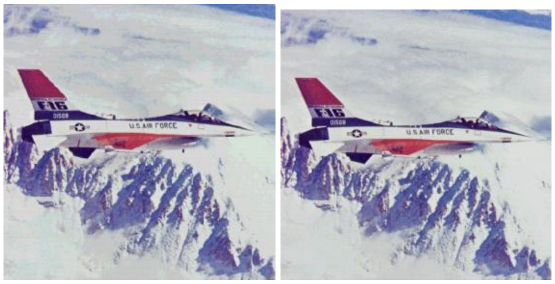
LSB LUCAS + PRNG

Fig.14. EMBEDDING TECHNIQUES PERFORMANCE (2ND BIT)



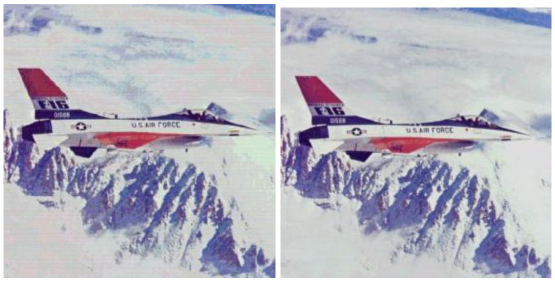
LSB LUCAS + PRNG

Fig.15. EMBEDDING TECHNIQUES PERFORMANCE (3RD BIT)



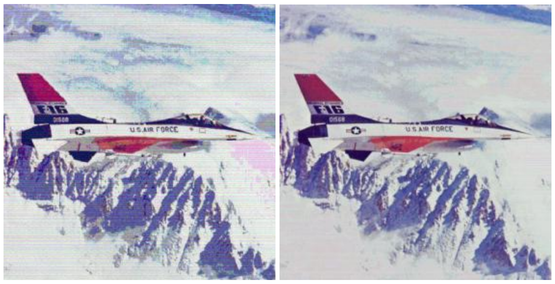
LSB LUCAS+ PRNG

Fig.16. EMBEDDING TECHNIQUES PERFORMANCE (4TH BIT)



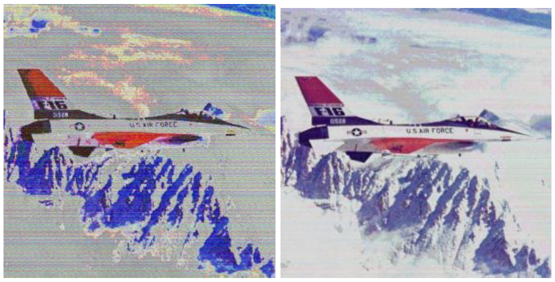
LSB LUCAS + PRNG

Fig.17. EMBEDDING TECHNIQUES PERFORMANCE (5TH BIT)



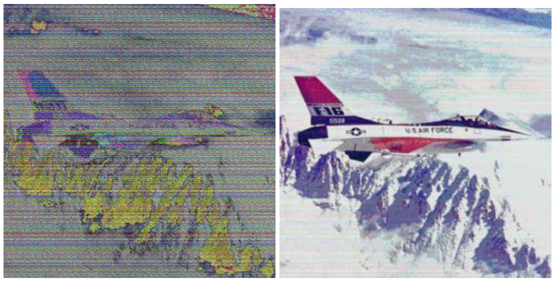
LSB LUCAS + PRNG

Fig.18. EMBEDDING TECHNIQUES PERFORMANCE (6TH BIT)



LSB LUCAS + PRNG

Fig.19. EMBEDDING TECHNIQUES PERFORMANCE (7TH BIT)



LSB LUCAS + PRNG

Fig.20. EMBEDDING TECHNIQUES PERFORMANCE (8TH BIT)

TABLE VIII. EMBEDDING TECHNIQUES PERFORMANCE (PSNR)

Bit plane	Peak Signal to Noise Ratio (PSNR)	
	LSB	Lucas + PRNG
1st	55.4721	53.5292
2nd	51.8705	54.0239
3rd	43.9574	51.1686
4th	36.5404	44.9436
5th	29.5467	40.1102
6th	23.4324	34.9022
7th	15.5457	30.1889
8th	9.2364	25.6393

V. CONCLUSIONS

Steganography is used to communicate an important data in a way that would not be noticeable by others. The least significant bit (LSB) is the most widely used technique for data hiding. The LSB process is simple but not secure. Also, using higher bit plane for hiding data would degrade the covered image's quality. In this paper, we are proposing a novel data hiding method based on Lucas number system. We use Lucas number system to decompose the images' pixels values to allow using higher bit plane for embedding without degrading the image's quality. Also, we enhanced the data security by using Pseudo Random Number Generators for selecting the image's pixels, colors and bits used for embedding secret data. Moreover, PRNGs are used to select the secret data bytes and bits to be embedded. The performance of the LSB and the proposed Lucas based method are evaluated by computing the Peak Signal to Noise Ratio (PSNR), where the proposed method achieved better performance than the LSB regarding the image quality and data security.

REFERENCES

[1] S. Katzenbeisser, F.A.P. Petitcolas, Information Hiding Techniques for Steganography and Digital Watermarking, Artech House, Norwood, MA, 2000.

[2] W. Bender, D. Gruhl, N. Morimoto, A. Lu, ―Techniques for data hiding‖ IBM Syst. J. 35 (3&4) (1996) 313–336.

[3] Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography, 2Nd Ed. ISBN: 978-0123725851

[4] Guorong Xuan, Yun Q. Shi, Zhicheng Ni, "Reversible data hiding using integer wavelet transform and companding technique," IWDW04, Korea, October 2004.

[5] Mauro Barni, Franco Bartolini, Vito Cappellini,Alessandro Piva(1998), "A DCT-domain system for robust image watermarking", Signal Processing, Vol. 66 (1998), pp.357–372.

[6] Jessica Fridrich and Miroslav Goljan, "On estimation of secret message length in LSB steganography in spatial domain," in Security, Steganography, and Watermarking of Multimedia Contents VI, Proceedings of SPIE 5306, pp. 23-34, 2004.

[7] Y. Qiudong, X. Liu, "A new LSB matching steganographic method based on steganographic information table", IEEE International Conference on Intelligent Networks and Intelligent Systems, pp. 362-365, 2009.

[8] S.M.M. Karim, M.S. Rahman, M.I. Hossain, "A new approach for LSB based image steganography using secret key", IEEE International

Conference on Computer and Information Technology, pp. 286-191, 2011.

[9] C.H. Yang, C.Y. Weng, S.J. Wang, H.M. Sun, “Adaptive data hiding in edge areas of images with spatial LSB domain systems”, IEEE Transactions on Information Forensics and Security, Vol. 3, pp. 488-497, 2008.

[10] K. Ghazanfari, S. Ghaemmaghami, S.R. Khosravi, “LSB++: An improvement to LSB+ steganography”, IEEE Region 10 Conference: Tencon 2011, pp. 364-368,2011.

[11] J. R. Hernandez, J. M. Rodr´ıguez, and F. P´erez-Gonz´alez, “Improving the performance of spatial watermarking of images using channel coding,” Signal Process. 80(7), pp. 1261–1279, 2000.

[12] Nabin Ghoshal , Jyostna Kumar Mandal, “A Novel Technique for Image Authentication in Frequency Domain using Discrete Fourier Transformation Technique”, Malaysian Journal of Computer Science, ISSN 0127-9094, Vol. 21, No. 1, pp. 24-32, 2008.

[13] F. Battisti, K. Egiazarian, M. Carli, and A.Neri, “Data hiding based on Fibonacci-Haar transform,” in Mobile Multimedia/Image Processing for Military and Security Applications, SPIE Defense and Security, Vol. 6579, May 2007.

[14] Chi-Kwong Chan and L. M. Cheng, “Hiding data in images by simple LSB substitution,” Pattern Recognition, pp. 469–474, Mar. 2004.

[15] Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography, 2Nd Ed. ISBN: 978-0123725851

[16] Swanson, M. D., Kobayashi, M., Tewfik, A. H.: Multimedia Data-Embedding and watermarking Technologies, Proc. IEEE, vol. 86, 1064 – 1087, 1998

[17] N. Johnson, Digital Watermarking and Steganography: Fundamentals and Techniques , The Computer Journal. (2009)

[18] G P. Ribenboim, My Numbers, My Friends, Springer, 2000, ISBN 0-38798911-0

[19] L. E. Dickson, "Recurring Series; Lucas' Un, Vn," History of the Theory of Numbers: Divisibility and Primality, Dover Publications, New York, Vol. 1, 2005, pp. 393-411.

[20] Brown, J. L. Jr. "Zeckendorfs Theorem and Some Applications", Fib. Quart. 2, 16 3-168, 1964.

[21] Phillips G.M., “Zeckendorf representation”, in Hazewinkel, Michiel, Encyclopaedia of Mathematics, Springer, ISBN 978-1556080104, Picione, 2001.

[22] L. T. Wang and E. J. McCluskey, “Linear feedback shift register design using cyclic codes,” IEEE Trans. Computer., vol. 37, pp. 1302-1306, Oct. 1988.

[23] A. Fuster and L. J. Garcia, “An efficient algorithm to generate binary sequences for cryptographic purposes,” Theoretical Computer Science, vol. 259, pp. 679-688, May 2001.

3D CAD model reconstruction of a human femur from MRI images

Mohammed RADOUANI
ENSAM, National Higher School of Engineering
Moulay Ismail University Meknès, Morocco

Benaissa EL FAHIME
ENSAM, National Higher School of Engineering
Moulay Ismail University
Meknès, Morocco

Youssef AOURA
ENSAM, National Higher School of Engineering
Moulay Ismail University
Meknès, Morocco

Latifa OUZIZI
ENSAM, National Higher School of Engineering
Moulay Ismail University
Meknès, Morocco

***Abstract*—Medical practice and life sciences take full advantage of progress in engineering disciplines, in particular the computer assisted placement technique in hip surgery. This paper describes the three dimensional model reconstruction of human femur from MRI images. The developed program enables to obtain digital shape of 3D femur recognized by all CAD software and allows an accurate placement of the femoral component. This technic provides precise measurement of implant alignment during hip resurfacing or total hip arthroplasty, thereby reducing the risk of component mal-positioning and femoral neck notching.**

***Keywords*—biomechanic; MRI imaging; 3D reconstructionl; femur**

I. INTRODUCTION

Collaborations between the medical and engineering communities in design and implantation of total joint replacements provide a means to restore function to a degraded human joint.

In case of hip surgery, Computer Assisted Surgery system needs the use of 3D CAD model of femur as reference shape to plan the virtual surgery. This model can be reconstructed using the images generated from scanners and enables observing the layers of the tissue shown in the images on a pixel by pixel level and in the context of the 3D structure of the human body.

Magnetic resonance imaging is a valuable imaging technique used primarily in medical settings to produce high quality images of the inside of the human body [1].

The clear anatomical images which can be produced in any plane, coupled with inherent safety for the patient and long-term cost effectiveness have promoted the use and availability of MRI around the world.

Automatic treatment techniques for visualization, interpretation and exploitation of these complex images have an enormous beneficial impact on clinical practice and research, by decreasing dramatically the manual effort which must otherwise be devoted [2]. In this work we develop a processing approach of MRI images for automatic reconstruction of a 3D model of the human femur.

II. MRI IMAGES TREATMENT OF THE FEMUR

A. Magnetic Resonance Imaging

Magnetic resonance imaging is based on the principles of nuclear magnetic resonance (NMR), a spectroscopic technic used by scientists to obtain microscopic chemical and physical information about molecules.

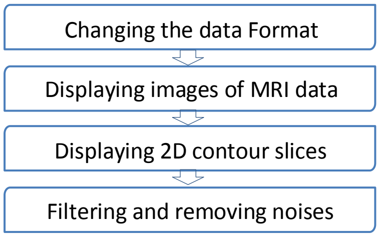
The technique was called magnetic resonance imaging rather than nuclear magnetic resonance imaging (NMRI) because of the negative connotations associated with the word nuclear in the late 1970's.

MRI started out as a tomographic imaging technique, that is it produced an image of the NMR signal in a thin slice through the human body. Each slice had a thickness. This form of imaging is in some respects equivalent to cutting off the anatomy above the slice and below the slice. The slice is said to be composed of several volume elements or voxels. The volume of a voxel is approximately 2 mm^3.

The magnetic resonance image is composed of several picture elements called pixels. The intensity of a pixel is proportional to the NMR signal intensity of the contents of the corresponding volume element or voxel of the object being imaged.

B. MRI images treatment

The process of 3D model reconstruction of femur is described by the following organizational chart:


Changing the data Format
Displaying images of MRI data
Displaying 2D contour slices
Filtering and removing noises

In first step, we have imported the MRI images into MATLAB workspace and saved it in a ".mat" file [3]. Then images are converted from grayscale to indexed ones. This procedure makes it possible to treat them as 4D matrixes. The set of 16 images are used in transversal plan (figure 1).

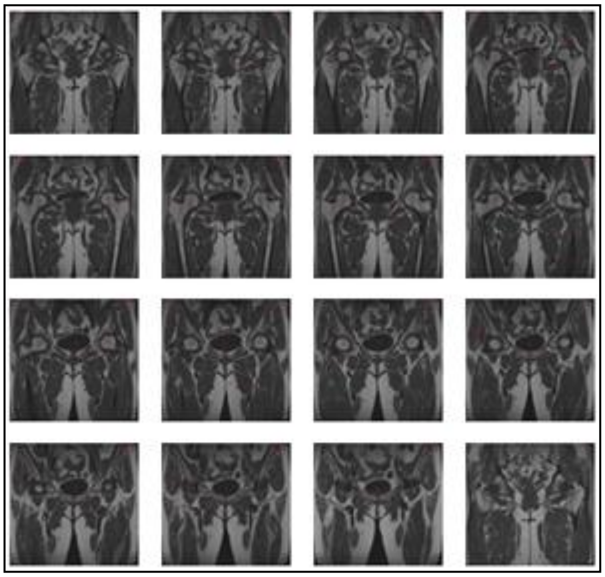


Fig.1. MRI images of femur in sagittal plane.

The whole set of 16 converted images is concatenated one on top of the other into a matrix using cat() function in Matlab software [4].

With this matrix an extra dimension derived from obsolete color maps is added. It can be eliminated using squeeze() function.

The Displaying 2D contour slices step consists on drawing contours in volume slice planes of all images. To eliminate artifacts, we designed the function to select the ROI (Region of Interest), and to convert the pixels out of ROI into the ones with the minimal value (figure 2).

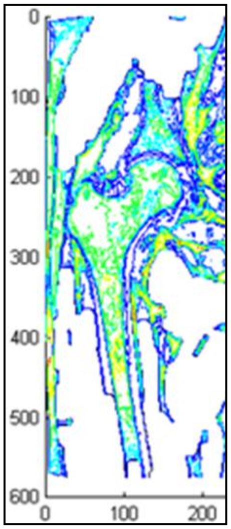


Fig.2. Contours detection of a single layer.

A function to filter the images obtained previously was programmed to eliminate residual noise. An example of the results of this operation is presented in figure 3.

The parameters which define the way of filtering are determined by the contrast and the color range of the acquired images.

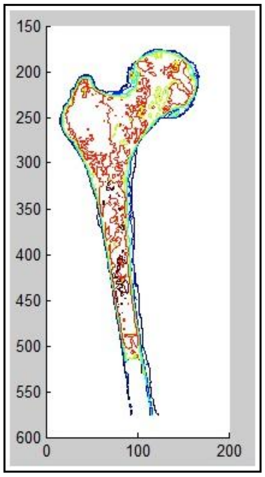


Fig.3. Bone shape extraction using the filtration function.

The code of the designed functions is as follows:

```
% load and convert images
    k=16
      for i=1:k
        Ai=imread('i.jpg');
        Ai=rgb2gray(Ai);
     end
    A = cat(4,A1,A2,A3,…,A16);
    A = squeeze(A);
% Displaying 2D contour slices
    cm = brighten(jet(length(map)),-.5);
    figure('Colormap',cm)
    contourslice(A,[],[],image_num)
    axis ij
    daspect([1,1,1])
    figure('Colormap',cm)
    Cr=contourslice(A,[],[ ],[1,2,3,…,16],8);
    view(3); axis  vis3d  tight
```

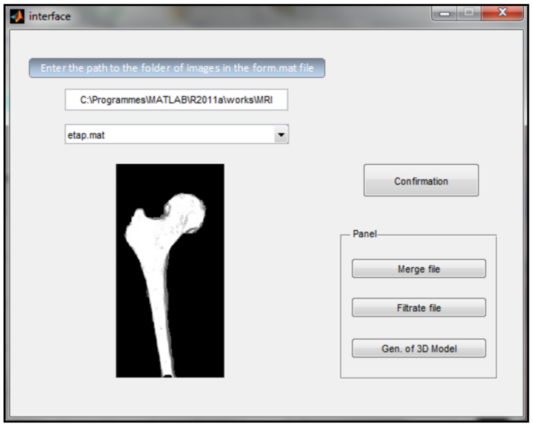
III. 3D Reconstruction of Femur

For The 3D automatic reconstruction of the femur a graphical user interface is developed in Matlab GUI environment [4]. This interface is presented in figure 4 and contains four buttons which initiate separate functions of the program:

The first one allows uploading and concatenating images into 3D variables, the whole set of images will be displayed as a MRI film.

Then the images will be filtered by clicking on the second button.

The third button will initiate the 3D reconstruction by creating and displaying contour slices with the same orientation and sizes as images (figure 5).



interface
Enter the path to the folder of images in the form.mat file
C:\Programmes\MATLAB\R2011a\works\MRI
etap.mat
Confirmation
Panel
Merge file
Filtrate file
Gen. of 3D Model

Fig.4. Bone shape extraction using the filtration function.

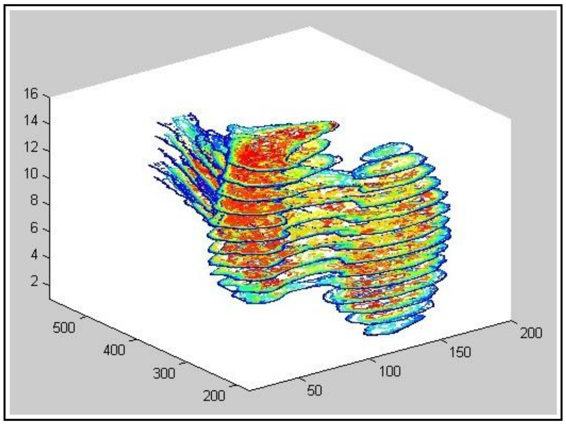


Fig.5. 3D display of contour slices.

At the end, the user has the option to enter the pixel by pixel coordinates of the planes which should divide the structure along the axes into a desired volume for observation.

Matlab editor for displaying 3D images has the ability of rotation and viewing the reconstruction from different angles.

The illumination of the display is set in such a way to show the best contrast in the most practical angles for viewing images (figure 6).

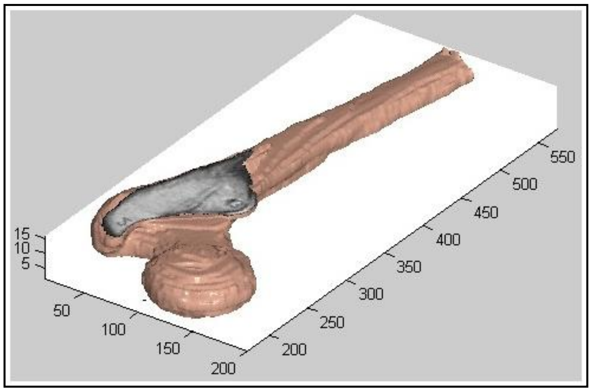


Fig.6. 3D model of the femur.

After having the 3D model of the bone, a STL file format is generated that can be used by all CAD software (figure 7).

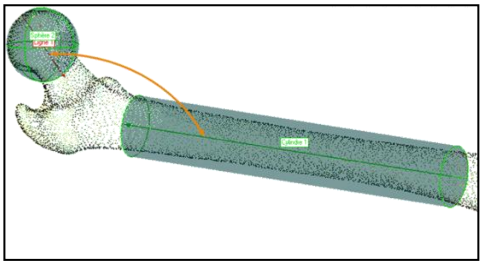


Fig.7. STL foramt of the femur

IV. Conclusion

The computer-assisted placement technique is an accurate and reproducible technique for hip surgery. Implants position may be better achieved for the navigated hips than by the traditional method.

In this study we proposed a method to generate a 3D CAD model of femur from MRI images. The developed technique allows the surgeon to make quickly the necessary decisions regarding the selection and positioning of the implants in 3D and reduce the risk of a misaligned component.

References

[1] W. Birkfellner, “Applied medical image processing : a basic course”. Boca Ratton, FL.: CRC Press, 2011.

[2] E. Ukwatta, J. Yuan, M. Rajchl, and A. Fenster, “Efficient global optimization based 3D carotid AB-LIB MRI segmentation by simultaneously evolving coupled surfaces”, Med Image Comput Comput Assist Interv, vol. 15, pp. 377-84, 2012.

[3] D. Zhou, W. K. Thompson, and G. Siegle, “MATLAB toolbox for functional connectivity”, Neuroimage, vol. 47, pp. 1590-607, Oct 1 2009.

[4] Matlab, Help, “sections: Visualizing MRI data: Volume Visualization Techniques (3-D Visualization) ”; Image Processing Toolbox.

Authors Profile

Mohammed RADOUANI is an associate professor at the National Higher School of Engineering (Crafts and Technologies, ENSAM Meknès - Moulay Ismail University, Morocco). He obtained his Ph.D. thesis in Mechanical Engineering from Prestigious training college for teachers and researchers in Technics (ENS of Cachan, University of Paris-south XI France, in 2003) and his Habilitation of supervising scientific research Dissertation from Faculty of Sciences-Meknès in 2009. His research work is dealing with specification and inspection of mechanical systems according to the ISO standards. He is also interested to products numerical engineering.

Youssef AOURA is an associate professor at the National Higher School of Engineering (ENSAM Meknès - Moulay Ismail University, Morocco). He obtained his Ph.D. thesis in Manufacturing Processes from the National Higher School of Engineering (ENSAM – France, in 2004). His research work is dealing with optimisation of new products and manufacturing processes.

Benaissa EL FAHIME obtained his Ph.D. thesis in Mechanical Engineering from the - Faculty of Science and Technology of Fez. His research activities concern the representation of mechanical tolerances in the Bond Graph models of mechatronic systems.

Latifa OUZIZI is an associate professor at the National Higher School of Engineering (ENSAM Meknès - Moulay Ismail University, Morocco). He obtained his Ph.D. thesis in Automatic and Productic from Metz University Metz France, in 2005. His research work is dealing with optimisation of new products and manufacturing processes.

Impact of other-cell interferences on downlink capacity in WCDMA Network

Fadoua Thami Alami
Abdelmalek Essaadi University
Morocco

Noura Aknin
Abdelmalek Essaadi University
Morocco

Ahmed El Moussaoui
Abdelmalek Essaadi University
Morocco

***Abstract*—Before the establishment of the UMTS network, operators are obliged to make the planning process to ensure a better quality of service (QOS) for mobile stations belonging to WCDMA cells. This process consists of estimating a set of parameters characterizing the radio cell in the downlink direction. Among them, we are interested in the Node B total required power P_{Tot}, and the maximum cell capacity, in the case of voice only and in the case of voice/video. To implement the effect of the other-cell interferences power, modeled as a fraction f_{DL} of the own-cell received power, on various radio parameters described previously, we focused our study on two different scenarios: the first is based on an isolated cell and the second, on multiple cells. In addition, when the WCDMA cell reaches its maximum capacity, the introduction of admission control algorithms is essential to maintain the QOS of the ongoing mobile stations. For this purpose, we have proposed an admission control algorithm, based both on the Node B total required power and the cell loading factor. This algorithm gives rigorous results compared to the existing ones in the literature.**

Keywords*—WCDMA; planning process; downlink; capacity estimation; other-cell interferences ;own-cell interferences; Node B total required power;admission control.

I. INTRODUCTION

The estimation of the WCDMA cell capacity is based on the signal to interferences ratio received at the level of an active mobile station in downlink and at the level of the Node B [1-2]. This ratio is expressed as a function of the required power of a mobile station *i* activating a service *j* and as a function of various problems harming the radio interface, such as: other-to-own cell interferences and thermal noise [1-4]. Thus, WCDMA network must allocate power by taking into account the required quality of service characterizing each service, and environmental conditions. This power is dynamically adjusted several times to preserve the energy per bit to noise spectral density ratio (E_b/N_0) constant [1-8].

In this work, we are interested in assessing and estimating the maximum downlink cell capacity in the case of a single-service and multi-services network, for two different scenarios: an isolated cell case, and a multiple-cells network case. The objective of this study is to highlight the effect of other-to-own-cell interference on WCDMA downlink capacity. In addition, network operators, look for controlling the admission of new mobiles stations, to prevent overloading status and fight against the degradation of link quality of communications already established. In this sense, we used two admission control algorithms to propose another, more efficient and rigorous, which takes into account two parameters: the maximum cell loading factor and Node B total available power P_{max}.

II. NODE B TOTAL POWER ESTIMATION IN A MULTI-SERVICE NETWORK

A. Link quality equation in a multiple cell and in an isolated cell cases

The estimation of the maximum cell capacity and Node B total required power is based on the link quality equation E_b/N_0 required for user *i* activating a certain service *j* (voice, video, web browsing, etc) in a cell *m*. This equation takes into account various radio interface problems: multipath propagation, neighboring cell effect, path loss, thermal noise equipment etc…

The expression of E_b/N_0 equation for a user *i* (i=1…$N_{user(j)}^{(m)}$) activating a service *j* ($j = 1...k$) in the cell *m*, in the case of a multiple cells case, can be written as:

$$E_b/N_0)_{ij}^{(m)} = \frac{W p_{ij}^{(m)}/L_{ij}^{(m)}}{R_{ij}^{(m)}((1-\alpha_{ij}^{(m)})(P_{Tot}^{(m)} - p_{ij}^{(m)})/L_{ij}^{(m)} + P_{Tot}^{(m)}\sum_{n=1,n\neq m}^{M} 1/L_{ijm}^{(n)} + P_N)} \qquad (1)$$

Where: P_{Tot}: Total downlink transmission power in the cell *m*; *M*: Number of cells in the network; P_N: Thermal noise power of the mobile; $R_{ij}^{(m)}$: The bit rate of a user *i* activating a service *j* in a cell *m* ; *W* : The chip rate ; $P_{ij}^{(m)}$: Transmission power required for a user *i* activating a service *j* in a cell *m*; $L_{ij}^{(m)}$: Pathloss between the Node B (of the cell *m*) and a user *i* activating a service *j* in the cell *m*; $L_{ijm}^{(n)}$: Pathloss between a Node B (of the cell *n*) and a user *i* activating a service *j* in the cell *m*; $\alpha_{ij}^{(m)}$: orthogonality factor of a user *i* activating a service *j* in the cell *m*. It depends on the multipath.

It is assumed that the total transmit power of different Node Bs is equal.

In the case of an isolated cell, the neighboring cells effect is absent. Then, the equation (1) should be modified as follows:

$$E_b/N_0)_{ij}^{(m)} = \frac{W p_{ij}^{(m)}/L_{ij}^{(m)}}{R_{ij}^{(m)}((1-\alpha_{ij}^{(m)})(P_{Tot}^{(m)} - p_{ij}^{(m)})/L_{ij}^{(m)} + P_N)} \qquad (2)$$

B. Node B total power expression

From equation (1) the required power expression of a mobile station i activating a service j in a cell m is:

$$p_{ij}^{(m)} = \frac{(1-\alpha_{ij}^{(m)} + f_{DLij}^{(m)})P_{Tot}^{(m)}}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{ij}^{(m)})} + \frac{P_N L_{ij}^{(m)}}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{ij}^{(m)})} \qquad (3)$$

Where:

$v_{ij}^{(m)}$ is the activity factor, which is used to evaluate the average utilization of the radio resources for the various services (voice, video ...). And we have:

$$f_{DLij}^{(m)} = \frac{I_{other}}{I_{own}} = \sum_{n=1, n\neq m}^{M} \frac{L_{ij}^{(m)}}{L_{ijm}^{(n)}} \qquad (4)$$

is defined as the other-to-own-cell interferences ratio received at a user *i* activating a service *j* in a cell *m*.

Where:

- I_{own}: own-cell interferences, includes the total received power from the users connected to the cell.

- I_{other}: other-cell interferences, includes the transmissions in the adjacent cells operating at the same frequency.

In downlink, this ratio $(0 \leq f_{DL} \leq 1)$ depends mainly on the user geographical position as well as on the neighboring Node Bs power. Thus it's different for each user. If the user is located on the cell edge, the f_{DL} value is high. On the contrary, if the user is close to his serving cell, f_{DL} value is small.

To simplify equation (3), we put:

$$X_{ij}^{(m)} = \frac{(1-\alpha_{ij}^{(m)} + f_{DLij}^{(m)})}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{ij}^{(m)})}, Y_{ij}^{(m)} = \frac{L_{ij}^{(m)}}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{ij}^{(m)})} \qquad (5)$$

And we consider that:

$$P_{Tot} = \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} p_{ij}^{(m)} \qquad (6)$$

Consequently:

$$P_{Tot} = \frac{P_N \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} Y_{ij}^{(m)}}{1 - \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} X_{ij}^{(m)}} \qquad (7)$$

Equation (3) can be rearranged as:

$$p_{ij}^{(m)} = (X_{ij}^{(m)} + Y_{ij}^{(m)} \frac{1 - \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} X_{ij}^{(m)}}{\sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} Y_{ij}^{(m)}}) P_{Tot} \qquad (8)$$

From equation (8), we conclude that the power consumed by the mobile station is a fraction of the Node B total available power, where:

$$k_{ij}^{(m)} = X_{ij}^{(m)} + Y_{ij}^{(m)} \frac{1 - \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} X_{ij}^{(m)}}{\sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} Y_{ij}^{(m)}} \qquad (9)$$

Is the amount of radio resources to be allocated to mobile station. Therefore, and by considering the user in an average position, the Node B total required power in a multi-service network becomes:

$$P_{Tot} = \frac{P_N L_{(moy)} \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \frac{1}{\frac{W}{(E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}}{1 - \left(1-\alpha_{(moy)} + f_{DL(moy)}\right) \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \left(\frac{1}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})} \right)} \qquad (10)$$

Where:

- $\alpha_{(moy)}$ is the average orthogonality factor,
- $f_{DL(moy)}$ is the average other-to-own-cell interferences ratio,
- $L_{(moy)}$ is the average pathloss in the cell.

In the case of an isolated cell, the other-to-own-cell interferences ratio f_{DL} is equal to zero. Consequently, equation (10) can be rewritten as: (5)

$$P_{Tot} = \frac{P_N L_{(moy)} \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \frac{1}{\frac{W}{(E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}}{1 - \left(1-\alpha_{(moy)}\right) \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \left(\frac{1}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})} \right)} \qquad (11)$$

C. WCDMA cell loading estimation

Equation (10) can be rearranged as:

$$P_{Tot} = \frac{P_N L_{(moy)} \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \frac{1}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}}{1 - \eta_{DL}} \qquad (12)$$

Where:

$$\eta_{DL} = \left(1-\alpha_{(moy)} + f_{DL(moy)}\right) \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \left(\frac{1}{\frac{W}{E_b/N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})} \right) \qquad (13)$$

η_{DL} is the cell loading factor. It increases with the number of users in the cell. Generally, we take $0 \leq \eta_{DL} < 1$ in order to maintain the system stability.

When $\eta_{DL} = 1$, the Node B total required power tends towards infinity and the system reaches its pole capacity. Thus, in the case of a voice service only, we obtain:

$$N_{user\_pôle}^{(m)} = \left(\frac{\frac{W}{E_b / N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}{1-\alpha_{(moy)} + f_{DL(moy)}} \right) \quad (14)$$

Therefore the maximum number of users should satisfy the following inequality:

$$N_{user\_\max}^{(m)} < \left(\frac{\frac{W}{E_b / N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}{1-\alpha_{(moy)} + f_{DL(moy)}} \right) \quad (15)$$

In the case of two services, the cell loading factor expression is the sum of cell loading factor generated by each service [6, 7].Thus:

$$\eta_{DL} = \sum_{j=1}^{k=2} \eta_{DL_j} \quad (16)$$

Where η_{DL_j} is the loading factor of a cell activate a service *j* and *k* is the number of services. In the case of an isolated cell, the cell loading factor can be rewritten as:

$$\eta_{DL} = \left(1-\alpha_{(moy)}\right) \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \left(\frac{1}{\frac{W}{E_b / N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})} \right) \quad (17)$$

D. Capacity and coverage tradeoff

From equation (12), the maximum path loss in downlink can be defined as:

$$L_{\max} = \frac{P_{\max}}{P_N \sum_{j=1}^{k} \sum_{i=1}^{N_{user(j)}^{(m)}} \frac{1}{\frac{W}{(E_b / N_0)_{ij}^{(m)} R_{ij}^{(m)} v_{ij}^{(m)}} + (1-\alpha_{(moy)})}} (1-\eta_{UL}) \quad (18)$$

E. Admission Control

When the system operates at nearly the maximum capacity, admitting a new user may affect the stability of the system.

Thus, admission control AC is crucial for balancing between the quality of service requirement of the new user and for the ongoing connections, and consequently preventing the system from an outage situation due to overloading. In this paper, we present two admission control algorithms presented in the literature.

The first is based on Node B total power and the second is based on cell loading factor [9-11]. We proposed another based on both the last two parameters.

Admission control strategies based on power estimation, must take into account whether the Node B has enough power to ensure the quality of service requirements for the new requests and for connections already established.

When a new call *i* arrives, the AC algorithm computes the minimum required transmission power for this call, and verifies the power constraint given by:

$$\sum_{a=1}^{n} k_{ij}^{(m)} P_{Tot} \leq P_{\max} \quad (19)$$

Where n is the new number of users if accepting the new call *i*. If the constraint is satisfied, the AC algorithm accepts the new call and allocates power $p_{ij}^{(m)}$ to the latter. Otherwise the call is blocked**.**

The second admission control algorithm is based on the maximum cell loading factor threshold. This parameter determines the admission of new calls in the network. When the cell loading factor of the Node B is below this threshold, a new call is accepted in the network. Above this threshold a new call is blocked at admission [11,12].

The following equation shows the constraint that must be verified:

$$\eta_{DL} + \Delta\eta \leq \eta_{DL\_\max} \quad (20)$$

Where η_{DL} is the current cell loading factor, $\Delta\eta$ the is the increase cell loading factor due to the new call and $\eta_{DL\_\max}$ is the maximum cell loading factor threshold setting by the operator at the dimensioning phase.

We propose an admission control algorithm which is based at the same time on the maximum cell loading factor and on the total available power at the Node B.

The following diagram illustrates this algorithm:

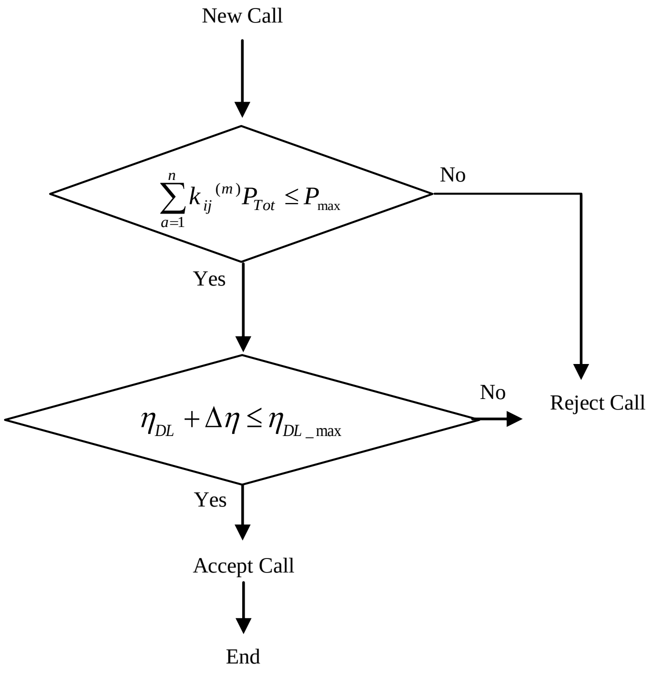

New Call
$\sum_{a=1}^{n} k_{ij}^{(m)} P_{Tot} \leq P_{\max}$
No
Yes
Reject Call
$\eta_{DL} + \Delta\eta \leq \eta_{DL\_\max}$
No
Yes
Accept Call
End

Fig.1. Admission control algorithm.

III. Simulation Parameters

The simulation parameters correspond to a macro-cellular network using omnidirectional antennas and Okumura-Hata propagation model which is written as:

$$L=137+35{,}2\log 10\ (d),\ d \text{ is in (km)} \qquad (21)$$

Where d is the distance separating the mobile and the serving Node B.

The average pathloss value was 133 dB and the maximal Node B power P_{max} was 20 Watt (43 dBm). The value of W is 3,84Mcp, the average orthogonality factor is 0,6 and the thermal noise is -100 dBm.

Figures 2 and 3 show the variation of the Node B total required power as a function of the number of users in a cell activating only voice service (R=12,2 kbit/s) and in a cell activating voice/video(R=66,5 kbit/s), respectively, for two different scenarios.

The target ratio E_b/N_0 for the voice service and video service are 5,5 dB and 6,5 dB respectively. The activity factor for voice and video are 0,67 and 1 respectively.

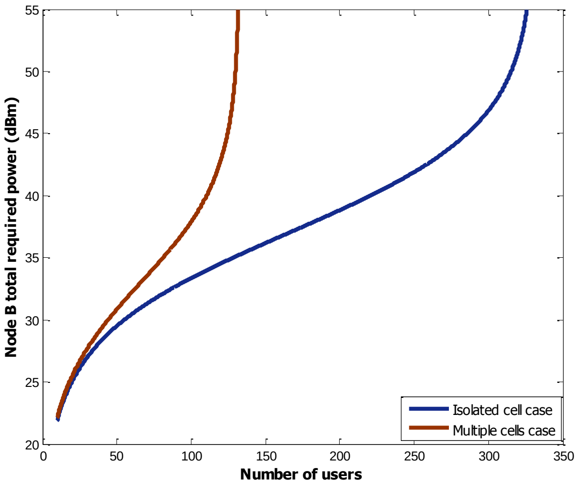


Fig.2. Node B total required power for voice service in the case of two scenarios.

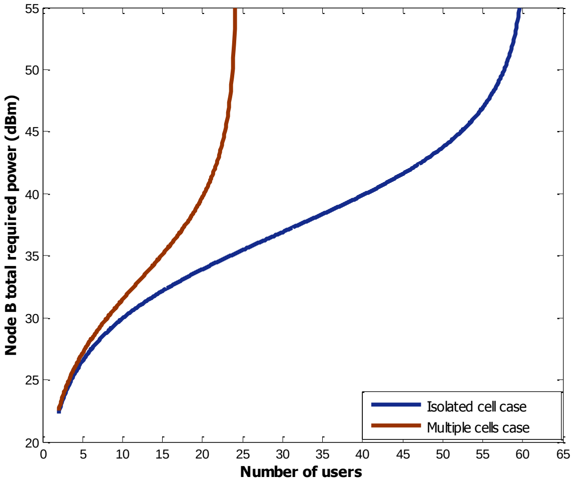


Fig.3. Node B total required power for voice/video services in the case of two scenarios.

From these figures, we note that the maximum cell capacity supported in an isolated cell case is more important than that supported in a multiple cells case. This is mainly due to the high-level of interference caused by the neighboring cells load in the last case, which lead to higher power consumption to ensure the target E_b/N_0. Therefore, the cell becomes loaded for a very small number of users.

In addition, the Node B power consumed by mobile stations in the case of a cell activating voice/video is much greater than that consumed in the case of a cell activating voice.

Figures 4 and 5 show the variation of the Node B total required power as a function of the number of users in a cell activating only voice service and in a cell activating voice/video respectively, for different values of f_{DL}.

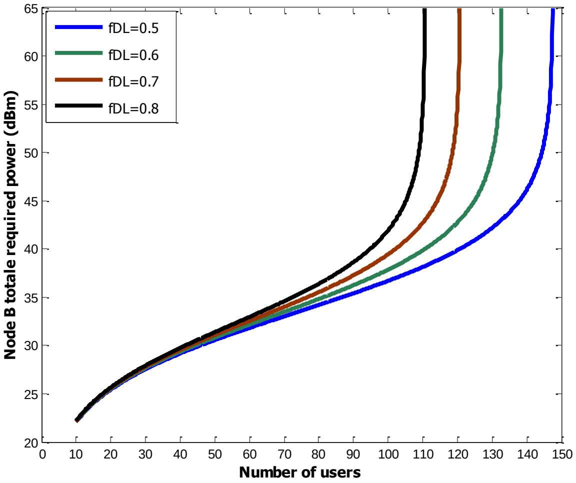


Fig.4. Node B total required power for voice service in the case of different values of f_{DL}.

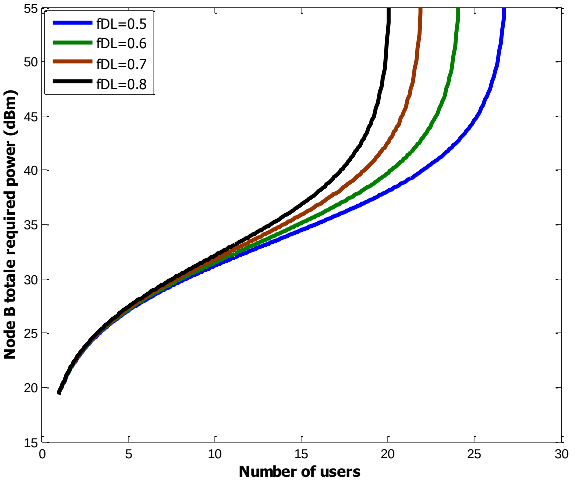


Fig.5. Node B total required power for voice/video services in the case of different values of f_{DL}.

According to figure 4, we notice that the Node B total required power increases with the number of active users in the cell, for a given value of f_{DL}.

By setting up the number of users and varying factor f_{DL}, we find that the value of the Node B total required power increases with the interference factor fDL. That is due to the increase of interference level in the cell, affecting the useful signal.

From figure 5, we note first that, the need of Node B total required power according to the number of active users in the cell increases in an important manner in comparison with the one in a cell activating only one service as shown in figure 4. Therefore, the Node B becomes saturated for a very small number of users since the total required power exceeds the maximum transmission power P_{max} available at the Node B. This means that the Node B cannot serve all the mobile stations with the required quality of service. On the other hand, the maximum capacity of a network with two services is much lower than the one of a single service network. This capacity decreases with increasing f_{DL} factor. For f_{DL}=0.5, the maximum number of users supported by the Node B is about 24 users, while for f_{DL} = 0.8 the maximum number supported is about 18 users.

The following figure shows the variation of the cell loading factor as a function of the number of users activating only voice service, for two types of scenarios.

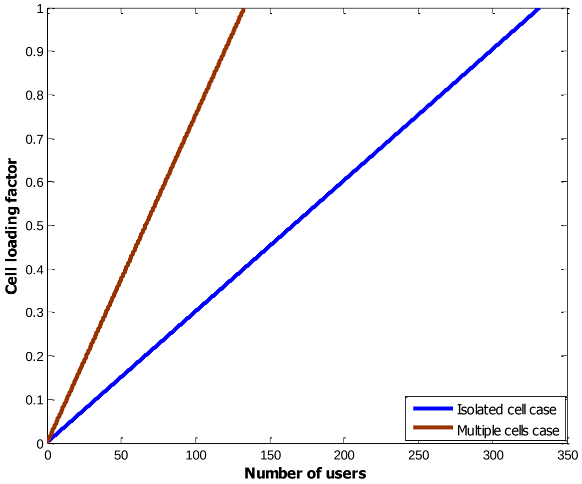


Fig.6. Cell loading factor as a function of the number of users activating voice service for two different scenarios.

According to this figure, we note that in the case of an isolated cell, the pole capacity is reached for approximately 331 users, whereas in the case of multiple cells, the pole capacity is reached for only 132 users approximately. This decrease in capacity is explained by the other-cell interferences that negatively affect the quality of service of active users in the cell of study. Therefore, users needing more power to meet their E_b/N_0 target value. Consequently, the maximum number of users that can support the cell decreases.

During the dimensioning phase, an operator may set, according to his needs, the value of the maximum cell loading factor $\eta_{DL\_\max}$. This value must be always strictly less than the pole cell loading factor.

In the case of two services: voice and video, the variation of cell loading factor for two different scenarios is presented in the following figure:

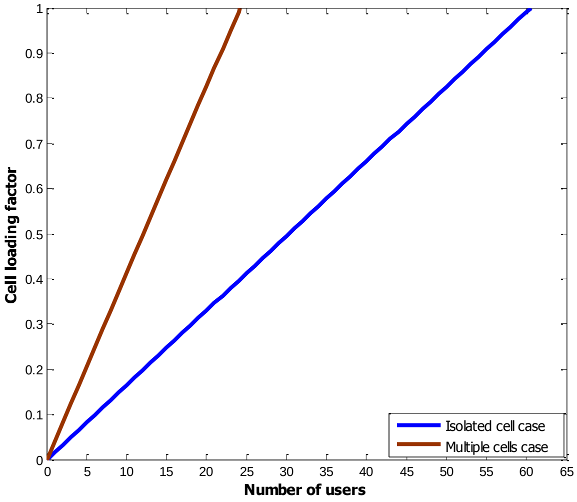


Fig.7. Cell loading factor as a function of the number of users activating voice/video services for two different scenarios.

From this figure we note that in the case of an isolated cell, the pole capacity is reached for approximately 60 users, while, in the case of multiple cells, the pole capacity is reached for only 24 users approximately. In addition, and according to figure 6, the pole capacity in both scenarios, is more important than in the case of two services: voice and video. This is due to the increase in power to devote to mobile stations in the latter case. Therefore, the maximum number of users that can support the cell decreases.

The following figure illustrates the variation of the transmitted power devoted to a given mobile station activating video service, according to the distance that separates it from its Node B. The coverage radius is taken equal to 2km.

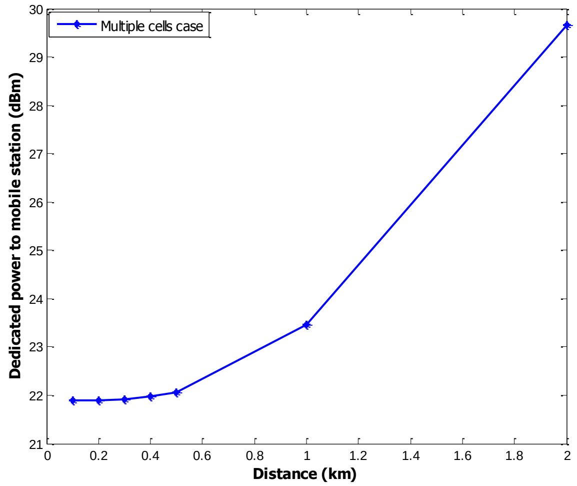


Fig.8. Dedicated power to mobile station according to distance in the case of multiple cells case.

From this figure and for distances less than 0.6 km, one can observe that the transmitted power dedicated to the mobile station as a function of the distance which separates it from its serving Node B, is almost constant. This is due mainly to the low pahloss existing for short distances, which means that the second term of the equation (3) is negligible compared to the first which does not depend on the pathloss.

For high distances, more the mobile station approaches the periphery of the cell; we observe that the transmitted power devoted to it increases significantly. This is due to the large distance between the mobile station and its serving Node B, which means that the second term of the equation (3) becomes

non negligible in long distances

Another cause of this increase in power is the high level of other-cell interferences existing in large distances which affects the quality of service of the existing mobile station on the cell periphery.

Figure 9 shows the blocking probability according to the number of new requests in the case of two scenarios by using the AC based on the Node B total power.

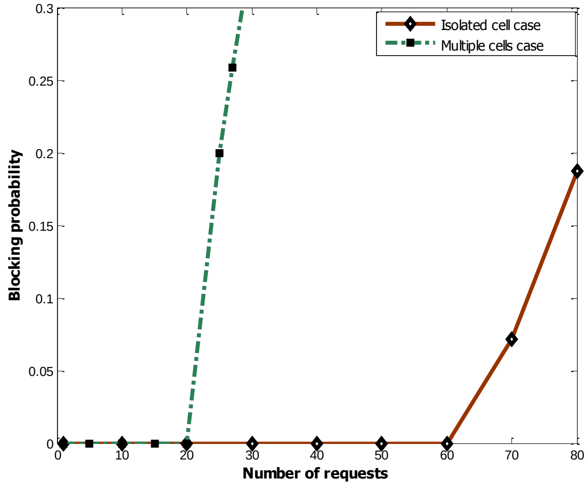


Fig.9. Blocking probability as a function of voice requests in the case of two scenarios.

From figure 9 we notice that the blocking probability in the case of an isolated cell, where there is an absence of other-cell interferences, is much smaller than that presented in the case of a network with multiple cells.

Effectively, in the case of a single cell, the number of requests simultaneously admitted is 60, while in the case of multiple cells; the cell can accept only 20 concurrent requests.

Figure 10 shows the blocking probability according to the number of new requests in the case of two scenarios by using the AC based on the cell loading factor.

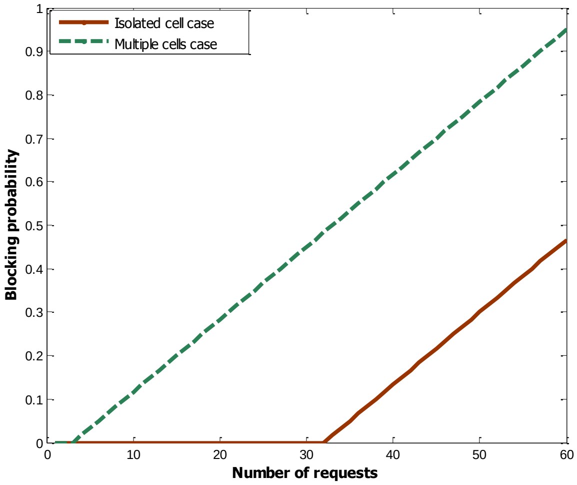


Fig.10. Blocking probability as a function of voice requests in the case of two scenarios.

From this figure, we note that, the blocking probability seen in the case of a single cell is much lower than in the case of multiple cells. This is mainly due to the absence of other-cell interferences in the first case, which makes the cell able to accept a large number of requests. According to this figure, the number of simultaneously accepted requests in the case of an isolated cell is approximately 32, while in the case of multiple cells the number of requests is approximately equal to 4.

Figure 11 shows the variation of the blocking probability according to the number of new requests in the case of the proposed algorithm:

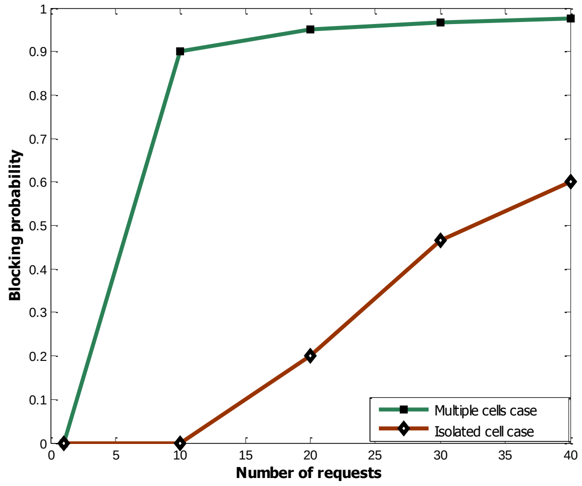


Fig.11. Blocking probability as a function of voice requests in the case of two scenarios.

From this figure we see that the blocking probability found with the proposed algorithm is greater than the one found by the application of power AC and cell loading factor AC for the two used scenarios. These results are more rigorous and more realistic.

IV. Conclusion

We have discussed in this paper the effect of other-to-own-cell interferences ratio f_{DL} on Node B total required power and maximum cell capacity, in downlink direction.

We found that more the other-cell interferences level is high, the more there is an increase in the first parameter and a decrease in the last one. In addition, in order to overcome the overloading situation, we present two admission control algorithms, the first is based on the Node B total power and the second is based on the cell loading factor. We have proposed a third which is based on the load and power at the same time. The application of this algorithm gives more realistic and rigorous results.

The study of the three AC algorithms is made for two different scenarios: An isolated cell and multiple cells cases. The results show that the effect of neighboring cells increase the blocking probability of new requests.

References

[1] Chie Dou, Yu-Hua Chang, Class-based downlink capacity estimation of a WCDMA network in a multiservice context, Computer Communications 28 (2005) 1443–1455, 13 January 2005.

[2] Kari Sipilä et al, Estimation of Capacity and Required Transmission Power of WCDMA Downlink Based on a Downlink Pole Equation, Vehicular Technology Conference 2000.

[3] Fadoua Thami Alami, Noura AKNIN, Ahmed El Moussaoui, Capacity estimation of multi-service network, Dimensioning and planification network, International Journal on Computer Science and Engineering, Vol. 3 Issue 3, 2011.

[4] Thrasivoulos (Sakis) Griparis, Tristan Lee, The capacity of WCDMA network: A case study, Bechtel Telecommunications Technical Journal, Vol. 3, No. 1, august 2005.

[5] Nicolas Enderlé, Xavier Lagrange, Analyse de la capacité descendante d'un système WCDMA, Actes du congrès DNAC, novembre 2001.

[6] Kimmo Hiltunen, Riccardo De Bernardi, WCDMA Downlink Capacity Estimation, Vehicular Technology Conference Proceedings. VTC 2000-Spring Tokyo. 2000 IEEE 51st, Volume 2, Issue, 2000 Page(s): 992 - 996 vol. 2. 2000.

[7] Jordi Pérez-Romero, Oriol Sallent, Ramon Agustı´, Radio Resource Management Strategies In UMTS, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2005.

[8] Soumaya Hamouda, Philippe Godlewski, Sami Tabbane, Downlink Capacity Estimation for UMTS Network: Impact Of Users'Position, Service Bit Rates And Cell Radius, Print ISBN: 978-9972-61-100-1, Issue Date: 11-14 Dec. 2005.

[9] J. Pérez-Romero, O. Sallent, D. Ruiz, R. Agustí, An Admission Control Algorithm to Manage High Bit Rate Static Users in W-CDMA, Dept. of Signal Theory and Communications, Universitat Politécnica de Catalunya.

[10] Gaurav Bansal, A. K. Chaturvedi†, and Vijay K. Bhargava, Distributed Admission Control for Power-Constrained Cellular Wireless Systems, Department of Electrical and Computer Engineering University of British Columbia, Vancouver, Canada, Print ISBN: 1-4244-1020-7, E-ISBN: 1-4244-1021-5, Issue Date: 22-26 April 2007.

[11] Zdeněk RŮŽIČKA, Stanislav HANUS, Admission Control And Load Control In UMTS Network, Institute of Radio Electronics, Brno University of Technology Purkyňova 118, 612 00 Brno, Czech Republic, contract no. 102/03/H109 and no. 102/04/2080.

[12] GUSTAVO AZZOLIN DE CARVALHO PIRES, Multi-Cell Admission Control for WCDMA Networks, Master of Science Thesis, Stockholm, Sweden 2006.

Hierarchical Low Power Consumption Technique with Location Information for Sensor Networks

Susumu Matsumae
Graduate School of Science and Engineering
Saga University
Saga 840-8502, Japan

Fukuhito Ooshita
Graduate School of Information Science and Technology
Osaka University
Osaka 565-0871, Japan

Abstract—**In the wireless sensor networks composed of battery-powered sensor nodes, one of the main issues is how to save power consumption at each node. The usual approach to this problem is to activate only necessary nodes (e.g., those nodes which compose a backbone network), and to put other nodes to sleep. One such algorithm using location information is GAF (Geographical Adaptive Fidelity), and the GAF is enhanced to HGAF (Hierarchical Geographical Adaptive Fidelity). In this paper, we show that we can further improve the energy efficiency of HGAF by modifying the manner of dividing sensor-field. We also provide a theoretical bound on this problem.**

Keywords—wireless sensor networks; geographical adaptive fidelity; energy conservation; network lifetime

I. INTRODUCTION

Wireless sensor networks have gained much attention in recent research and development. In the wireless sensor networks, battery-powered sensor nodes are placed on the observation area, and the sensed data is transmitted to the observer by multi-hop communication between nodes. Traditionally, the routing protocols for these networks have been evaluated in terms of packet loss rates, routing overhead, etc. However, since wireless sensor networks are usually deployed using battery-powered nodes, the optimization of routing protocol's energy consumption is also important [1], [5], [6].

The usual technique for designing an energy-efficient routing protocol is to activate only necessary nodes (e.g., those nodes which compose a backbone network) and to put other nodes to sleep [2], [3], [4], [7], [8]. Among these protocols, in this paper we focus on GAF (Geographical Adaptive Fidelity) [8] and its extended versions called HGAF (Hierarchical Geographical Adaptive Fidelity) and eHGAF (extended HGAF) [4].

In this paper, we show that we can improve the energy efficiency of eHGAF[4] by modifying the manner of dividing sensor field. In the GAF-based algorithms, the sensor field is partitioned by regions called cells, and the cell size affects the energy efficiency of the protocols. Table I summarizes the maximum cell sizes for the GAF-based methods. As shown in Table I, in this paper we successfully obtain the cell size of $\sqrt{3}R^2 \approx 1.73205R^2$, which is 73.205% larger than that of eHGAF [4] whose cell size is R^2.

In this paper, we also study an upper bound on the cell size, and prove the upper bound $\pi R^2 - \Delta$ where R is the radio range of each sensor node and $\Delta = \frac{4\pi - 3\sqrt{3}}{6}R^2$. Since this upper bound is approximately $1.91R^2$, our method which attains $\sqrt{3}R^2$ is fairly closed to the theoretical upper bound.

The rest of this paper is organized as follows. Section II explains the outline of GAF, HGAF, and eHGAF. Section III shows that we can further increases the energy efficiency of eHGAF. Section IV provides the upper bound on the cell size. And finally Section V offers concluding remarks.

TABLE I. MAXIMUM CELL SIZES FOR GAF-BASED PROTOCOLS

	Max cell size
GAF [8]	$\frac{1}{5}R^2$
HGAF [4]	$\frac{1}{2}R^2$
eHGAF [4]	R^2
eHGAF with triangle cells (this paper)	$\frac{3\sqrt{3}}{4}R^2$
eHGAF with two cell types (this paper)	$\sqrt{3}R^2$

Here, *R* is the radio range of each sensor node.

II. PRELIMINARIES

Throughout the paper, as in [4], [8], we assume that the radio range of each node is R and is unchanged during the operation, and that every node knows its own location information.

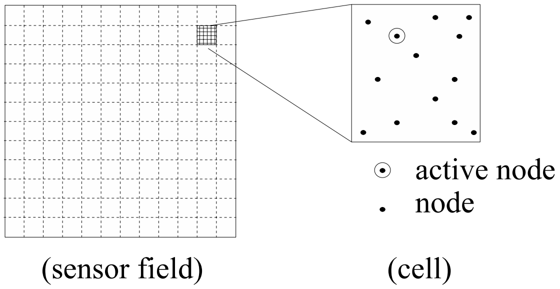


Fig.1. A sensor field divided by square cells

A. GAF (Geographical Adaptive Fidelity)

In GAF [8], the entire sensor field is divided into virtual sub-fields called *cells*. In each cell, a node called *active node* is chosen. These active nodes have the following two missions:

1) The active nodes compose a backbone network for inter-cell data transmissions. Every data across cell-boundaries is conveyed through this back- bone in multi-hop manner.

2) Each active node acts as a gateway node of its own cell. Every transmission across the cell-boundary is via the gateway node.

Each active node is not fixed, and is properly changed over by the other node in the same cell, according to the remaining amount of battery at the time. The election is dynamically performed by a leader-election algorithm described in [8]. The active nodes are steadily activated, while other nodes are activated only when necessary and are being asleep most of the time.

From the viewpoint of energy consumption, it is preferable to make the cell as large as possible [4]. This is because the larger a cell becomes, the smaller the total number of active nodes in the entire sensor field is. The cell size, however, has an upper bound, and we cannot make it larger without limitation. The upper bound is subject to the communication range of each sensor nodes and the following two requirements:

(Req. I) Any pair of active nodes can communicate with each other if their cells are adjacent.

(Req. II) Any active node can communicate with every other node within the cell.

The requirements (Req. I) and (Req. II) are necessary for assuring the missions (I) and (II) of active nodes, respectively.

In GAF, the sensor field is simply divided by square-shaped cells of the same size (see Fig. 1). Let the size of cell be $r \times r$. For GAF, the requirements (Req. I) and (Req. II) are respectively taken on concrete formulas as follows:

$$r^2 + (2r)^2 \leq R^2, \qquad \text{(Req. I-GAF)}$$

$$r^2 + r^2 \leq R^2. \qquad \text{(Req. II-GAF)}$$

The inequality (Req. I-GAF) is due to Fig. 2(a), and the inequality (Req. II-GAF) is due to Fig. 2(b). From these inequalities, we have

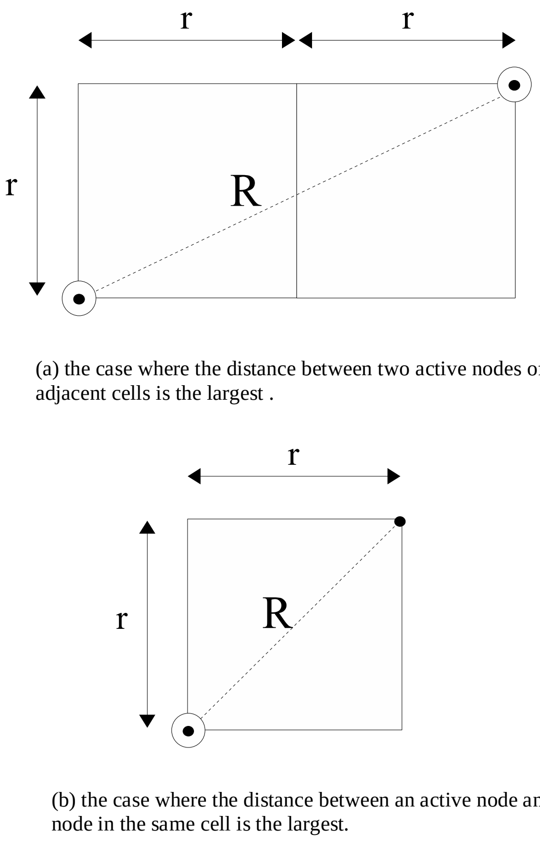
$$r \leq \frac{R}{\sqrt{5}}$$

and thus the following claim holds:

Claim 1 [8] In GAF, the cell size is bounded above by $\frac{R^2}{5}$. ■

B. HGAF (Hierarchical Geographical Adaptive Fidelity)

In HGAF [4], the cell size can be $\frac{R^2}{2}$ at largest, by relaxing the dominant condition (Req. I-GAF) of GAF.



(a) the case where the distance between two active nodes of adjacent cells is the largest .

(b) the case where the distance between an active node and a node in the same cell is the largest.

Fig.2. Examples supporting (Reg. I) and (Req. II) for GAF

The key idea is to avoid the extreme case illustrated in Fig. 2(a). In HGAF, each cell is further divided into smaller squares called *subcells*. A cell of size $r \times r$ is divided into subcells of size $d \times d$. For simplify the exposition, we consider only the case where r is divisible by d.

A subcell is called *active subcell* if it contains an active node of the cell. In HGAF, active subcells are maintained in the same position of the respective cells, and their positions are synchronously rotated. By this modification, as for (Req. I), we have only to consider the case illustrated in Fig. 3. As for (Req.

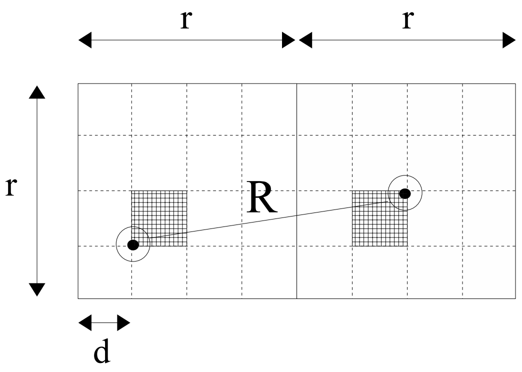


Fig.3. An example supporting (Reg. I) for HGAF. Here, the distance between two active nodes of adjacent cells is the largest.

II), the inequality is the same as that of GAF. Hence, we have

$$d^2 + (r+d)^2 \le R^2, \qquad \text{(Req. I-HGAF)}$$

$$r^2 + r^2 \le R^2. \qquad \text{(Req. II-HGAF)}$$

From (Req. I-HGAF), we have

$$r \le \sqrt{R^2 - d^2} - d,$$

and r can be R at largest when we let d be infinitesimal (i.e., the partition of each cell into subcells is infinitely fine-grained). Hence, the constraint (Req. II-HGAF) is the dominant condition here, and thus the following claim holds:

Claim 2 [4] In HGAF, the cell size is bounded above by $\frac{R^2}{2}$. ■

C. eHGAF (extended HGAF)

In eHGAF [4], the cell size is improved and can be R^2 at largest. Here, the dominant constraint for HGAF is relaxed by keeping active subcells centered. For simplicity, we assume that r is divisible by d and that the quotient is an odd number.

To place an active subcell in the center of its cell, the cell-boundaries are synchronously slided properly. By this modification, for (Req. II), we have only to consider the case illustrated in Fig. 4. As for (Req. I), the inequality is the same as that of HGAF. Hence, we have

$$d^2 + (r+d)^2 \le R^2, \qquad \text{(Req. I-eHGAF)}$$

$$2\left(\frac{r+d}{2}\right)^2 \le R^2. \qquad \text{(Req. II-eHGAF)}$$

From (Req. I-eHGAF), r can be R at largest when we let d be infinitesimal. From (Req. II-eHGAF), r can be $\sqrt{2}R$ at largest for infinitesimal d. Hence, the constraint (Req. I-eHGAF) is the dominant condition here, and thus the following claim holds:

Claim 3 [4] In eHGAF, the cell size is bounded above by R^2. ■

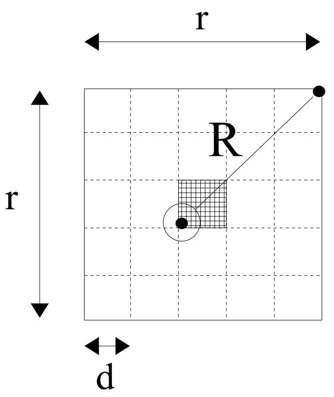


Fig.4. An example supporting (Req. II) for eHGAF. Here, the distance between an active node and a node in the same cell is the largest.

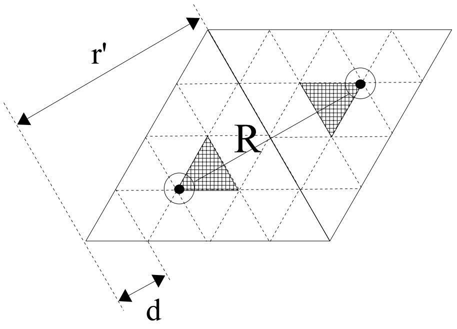


Fig.5. An example supporting (Req. I) for eHGAF with triangle cells. Here, the distance between two active nodes of adjacent cells is the largest.

III. PROPOSED METHOD

A. Cell Enlargement by Changing Cell Shape

In eHGAF with triangle cells, the upper bound on the cell size obtained for the standard eHGAF (i.e., eHGAF with square cells) is further improved. Actually, the cell size can be approximately $1.29904R^2$ at largest, which is 29.904% larger than that of eHGAF [4].

The main idea is to change the base-shape of each cell (and subcell) to triangle cells.

Previously in GAF, HGAF, and eHGAF, the cells are of square-shape. Although partitioning with squares is regular and natural, a plane can be tiled with other regular polygons such as regular triangle and regular hexagon.

In GAF, the cell size can be $\frac{1}{4\sqrt{3}}R^2 \approx 0.14434R^2$, if we use triangle cells, and be $\frac{3\sqrt{3}}{26}R^2 \approx 0.19985R^2$, if we use hexagon cells. That is, for GAF, we cannot improve the upper bound on the cell size even if we adopt triangle/hexagon cells.

However, in eHGAF, the upper bound on the cell size can be improved up to approximately $1.29904R^2$ if we use triangle cells. See Fig. 5. For simplify the exposition, we assume that r' is divisible by d and that the quotient is ($3c$+1) for some positive integer c.

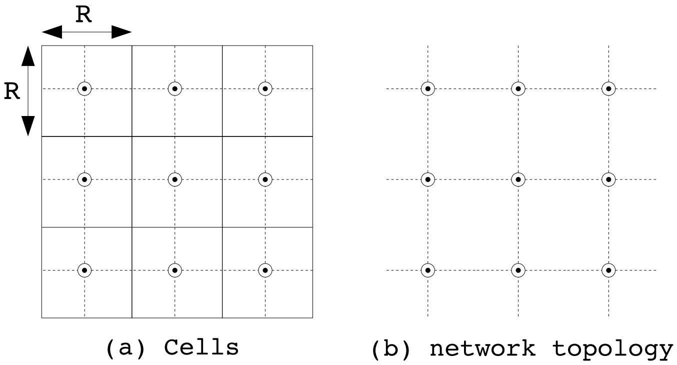


Fig.6. The cell partition, and the network graph of active nodes in eHGAF

This assumption assures that an active subcell can be located in the center (barycenter) of the triangle cell. Here, when *d* is infinitesimal, we can think that the active subcell can be seen as the point just positioned at the barycenter of the regular triangle. In such a case, it is easy to check that the conditions (Req. I) and (Req. II) become identical, and we have the following one inequality:

$$r' \leq \frac{3}{2}R \quad (d \text{ is infinitesimal}).$$

Since r' is the height of a regular triangle, the maximum size of triangle-shaped cell is calculated as

$$\frac{1}{2} \cdot \frac{2}{\sqrt{3}} r' \cdot r' = \frac{3\sqrt{3}}{4} R^2 \approx 1.29904R^2.$$

Hence, we obtain the following theorem:

Theorem 1 In eHGAF with triangle cells, the cell size is bounded above by $\frac{3\sqrt{3}}{4}R^2 \approx 1.29904R^2$. ■

B. Cell Enlargement by Reducing Edges

Next, we show that we can further improve the upper bound on the cell size to $\sqrt{3}R^2 \approx 1.73205R^2$, which is 33.333% larger than that of eHGAF with triangle cells.

In the preceding subsection, the upper bound of $\frac{3\sqrt{3}}{4}R^2 \approx 1.29904R^2$ is obtained by adopting triangle cells. Here in this section, we use a different approach and consider reducing edges of network graph of active nodes.

0) Relaxing (Req. I):

In eHGAF with square-cells, the cell size is bounded above by $R \times R$, and each active node is located around the center of its belonging cell. Due to (Req. I), any pair of active nodes must be capable of communicating with each other if their cells are adjacent, and hence the network graph of active nodes becomes a square mesh/lattice. See Fig. 6.

As long as we adhere (Req. I), we cannot break the upper bound of $R \times R$. However, if we loosen (Req. I) properly, we can improve the upper bound further. Here, for example, we consider a version of (Req. I) as follows:

(Req. I') Any pair of active nodes can communicate with each other if their cells are *horizontally* adjacent.

By such a relaxation, though the width of a cell cannot be made longer, the height of it can be $\sqrt{3}R$ at longest. Here, it should be noted that the cell of size $\sqrt{3}R \times R$ can be included in the circle with radius *R*. See Fig. 7.

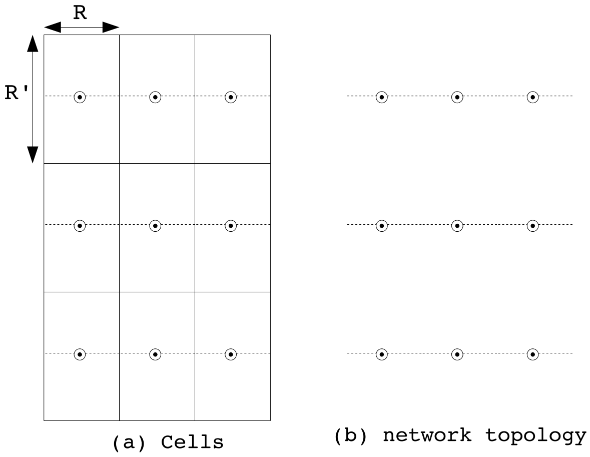


Fig.7. The cell partition of R′ × R where R′ = √3R, and the network graph of active nodes in eHGAF. Here we adopt (Req. I') instead of (Req. I).

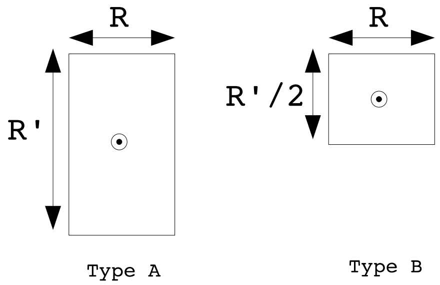


Fig.8. Two types of cells for eHGAF. Here, R′ = 3R.

1) Partition with Two Types of Cells:

To obtain connectivity of entire network, we consider using another type of cells as well. Here, we use the two types of cells shown in Fig. 8. The height of type B cell is half of that of type A cell. Since the height of type B is $\frac{\sqrt{3}}{2}R$ and is less than *R*, the active node located in the center of it can communicate with its counterpart of upper/lower adjacent cell of type B.

By using both type A and type B cells, we can partition a sensor field in such a way that the network graph of active nodes is connected. See Fig. 9 for an example. If we assume that the columns composed of type B cells are placed every *k* columns, the average size of cells is calculated as $\frac{\sqrt{3}kR^2}{k+1}$, which converges on $\sqrt{3}R^2 \approx 1.73205R^2$ when *k* is infinitely large. Here, it should be noted that even when we chose *k* = 3, the average size of cells already becomes $\frac{3\sqrt{3}}{4}R^2$, which is the same size obtained in eHGAF with triangle cells.

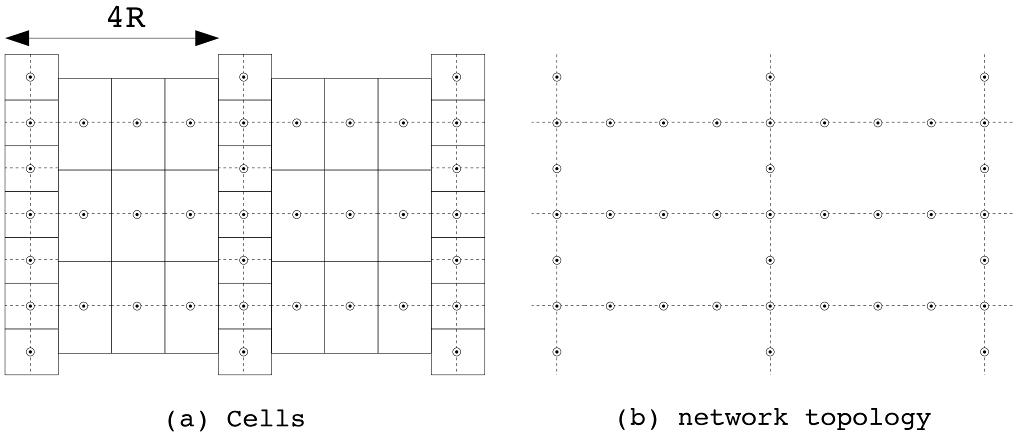


Fig.9. The cell partition with type A and type B cells, and the network graph of active nodes in eHGAF. Here, the columns composed of type B cells are placed every 4 columns.

Hence, we can state the following theorem:

Theorem 2 In eHGAF with square cells, the cell size can be $\sqrt{3}R^2 \approx 1.73205R^2$, at largest if we permit the existence of active nodes whose degree is less than 4. ■

IV. Upper Bound of Cell Size in eHGAF

In this section, we study a theoretical upper bound on the cell size for eHGAF. We show that the cell-size is asymptotically bounded above by $\pi R^2 - \Delta$ in average, where $\Delta = \frac{4\pi - 3\sqrt{3}}{6} R^2$. Here, we do not assume that the shape of sensor field nor that of cells.

To begin with, we introduce the following two propositions.

Proposition 1 If a sensor field consists of a single cell, the size of entire sensor field can be πR^2 at largest. ■

Proposition 2 If a sensor field consists of 2 cells, the size of entire sensor field can be $2\pi R^2 - \Delta$ at largest. ■

An example for Proposition 1 is a sensor field whose shape is a circle with radius *R*. An example for Proposition 2 is the one whose shape is the union of two circles such that the radius is *R* for both circles and that the distance between their centers is *R*.

By generalizing the above two propositions, we can prove the following lemma.

Lemma 1 If a sensor field consists of n cells, the size of entire sensor field can be $n\pi R^2 - (n-1)\Delta$ at largest. ■

The Lemma 1 can be proved by mathematical induction on *n*. The base case is due to Proposition 1 and 2. The inductive case can be proved by the following lemma:

Lemma 2 Let S_k be a sensor field composed of k cells. Construct S_{k-1} from S_k by the following 3 steps:

0) *choose any single cell C of S_k,*
1) *remove the region of C if it is covered only by the active node of C, and*
2) *migrate the region of C to the adjacent cell C' if it can be covered by the active node of C'.*

Then, the following inequality holds:

$$|S_k| - |S_{k-1}| \leq \pi R^2 - \Delta$$

where $|S|$ denotes the area of S. ■

Lemma 2 can be easily checked by the following observation. If an active node of a cell C_1 can communicate with its counterpart of adjacent cell C_2, then there exists an overlapped area for the circle with radius R whose center is the active node of C_1 and that whose center is the active node of C_2, and the size of that overlapped area has to be Δ at least.

Let $P(k)$ denote the following proposition:

If a sensor field consists of k cells, the size of entire sensor field can be $k\pi R^2 - (k-1)\Delta$ at largest.

For the proof of inductive case for Lemma 1, if we assume that $P(k)$ holds and that $P(k+1)$ does not, then we can derive a contradiction by Lemma 2.

From Lemma 1, the average cell size can be calculated as

$$\frac{n\pi R^2 - (n-1)\Delta}{n} = \pi R^2 - \frac{n-1}{n}\Delta.$$

Hence, we can derive the following theorem as follows.

Theorem 3 In eHGAF, the average cell size can be asymptotically $\pi R^2 - \Delta$ at largest. ■

V. Concluding Remarks

In this paper, we showed the following:

1) The cell size of eHGAF can be $\left(\frac{3\sqrt{3}}{4}\right)R^2 \approx 1.29904R^2$ *at largest if we use triangle cells.*

2) The cell size of eHGAF can be $\sqrt{3}R \approx 1.73205R^2$ *at largest if we permit the existence of active nodes whose degree is less than 4.*

3) The upper bound on the cell size of eHGAF is $\pi R^2 - \Delta$ *where* R *is the radio range of each sensor node and* $\Delta = \frac{4\pi-3\sqrt{3}}{6}R^2$.

As shown in Table I, since the previous result [4] can attain only R^2 at largest, our results successfully improve the upper bound on the cell-size for the GAF-based methods. Further, since the theoretically obtained upper bound, $\pi R^2 - \Delta$, is approximately $1.91R^2$, our method which attains $\sqrt{3}R^2$ is fairly closed to the theoretical bound.

Since the total number of active nodes in the entire sensor field inversely relate to the cell size, we can say that the energy use of our scheme is more efficient than the previous ones [4], [8]. Table II summarizes the estimated network lifetime for the GAF-based methods. Here, we assume that the network lifetime is proportional to the inverse of the number of active nodes in the entire network.

For future work, we will study whether we can improve the cell-size and/or the theoretical bound further.

TABLE II. NETWORK LIFETIME FOR GAF-BASED PROTOCOLS

	The network lifetime compared to the theoretical upper bound
GAF [8]	11%
HGAF [4]	26%
eHGAF [4]	52%
eHGAF with triangle cells (this paper)	68%
eHGAF with two cell types (this paper)	91%
theoretical upper bound	100%

ACKNOWLEDGMENT

This work was partly supported by the JSPS Grant-in-Aid for Young Scientists (B) (23700056).

REFERENCES

[1] J.-H. Chang and L. Tassiulas. Energy conserving routing in wireless adhoc networks. In INFOCOM (1), pages 22–31, 2000.

[2] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris. Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. In Mobile Computing and Networking, pages 85–96, 2001.

[3] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan. An application-specific protocol architecture for. IEEE Transactions on Wireless Communications, 1 (4), 2002, 2002.

[4] T. Inagaki and S. Ishihara. A proposal of a hierarchical power saving technique using location information for sensor networks. IPSJ SIG Technical Report (in Japanese), 2007(14), pages 1–8, 2007.

[5] D. Kim, J. Garcia-Luna-Aceves, K. Obraczka, J. Cano, and P. Manzoni. Power-aware routing based on the energy drain rate for mobile ad hoc networks. In the IEEE International Conference on Computer Communication and Networks, 2002.

[6] S. Singh, M. Woo, and C. S. Raghavendra. Power-aware routing in mobile ad hoc networks. In Mobile Computing and Networking, pages 181–190, 1998.

[7] Y. Xu, J. Heidemann, and D. Estrin. Adaptive energy-conserving routing for multihop ad hoc networks. Research Report 527, USC/Information Sciences Institute, 2000.

[8] Y. Xu, J. S. Heidemann, and D. Estrin. Geography-informed energy conservation for ad hoc routing. In Mobile Computing and Networking, pages 70–84, 2001

Translation of Pronominal Anaphora from English to Telugu Language

T. Suryakanthi
Research Scholar, Dept. of CSE
Lingaya's University
Faridabad, Haryana, India

Dr. S.V.A.V. Prasad
Dean of R&D
Lingaya`s University
Faridabad, Haryana, India

Dr. T. V. Prasad
Dean of Computing Sciences
Visvodaya Technical Academy
Kavali, Andhra Pradesh, India

***Abstract*—Discourses are linguistic structures above sentence level. Discourse is nothing but a coherent sequence of sentences. Discourse analysis is concerned with coherent processing of text segments larger than the sentence and this requires something more than just the interpretation of the individual sentences. A phenomenon that operates at discourse level includes cohesion. Text is cohesive if its elements link together. This linking can be either forward or backward. Pronominal referencing is one method for linking sentences. This paper presents the issues in translating pronominal references from English to Telugu language. This work handles resolution and generation of personal pronouns whose antecedents appear before the anaphora. An algorithm is developed for translation of pronominal references.**

Keywords*—GNP; Gender number person; SL: Source language; English; TL; target language: Telugu; S-singular; P-plural, M-masculine; F-feminine; N-neuter; VBD – past tense verb form; VBZ- 3rd person singular present verb form; VBP- non 3rd person singular present verb form; MD- Modal

I. INTRODUCTION

Bloomfield and Chomsky 1957 have defined that the sentence is the largest grammatical unit for language analysis. Halliday and Hasan (1976) threw light on the concepts like coherence and cohesion. Discourse analysis is concerned with coherent processing of text segments larger than the sentence and this requires something more than just the interpretation of the individual sentences. Machine translation refers to the task of translating text from one natural language to another with minimal human intervention.

The present machine translation system is a rule based machine translation system, where parallel grammar was developed for both source and target languages. Phrase structure grammar framework is used to develop the grammar rules for the languages. This system is able to translate text above sentence level. The text above sentence level also called the discourse text. Translation of discourse includes resolving references used in the sentences. This paper presents the resolution and evaluation of these anaphora problems in translating from English to Telugu language.

II. PRONOMINAL REFERENCE AND ANAPHORA

A grammatical term for pronoun, which refers back to another word or phrase, is called Anaphora. Halliday and Hasan defined anaphora as the cohesion which points back to some previous item [1]. The item which refers is called anaphor and the item which is referred is called the antecedent.

Ex: 1 Ram went to fruit market. He likes apples very much.

In the above sentence 'He' is a pronoun which refers to Ram in the previous sentence. Here 'He' is an anaphor and 'Ram' is an antecedent. This is the most common type of anaphor called the pronominal anaphora. Anaphora phenomenon has two processes, resolution and generation. 'Resolution' refers to the process of determining the antecedent of an anaphor; 'Generation' is the process of creating references over a discourse entity. This work handles resolution and generation of personal pronouns whose antecedents appear before the anaphora. cataphoric relations are not taken into account in this study. The translation of third person personal pronouns from English to Telugu language has been evaluated on unrestricted corpora. The precision achieved in translating personal pronouns is above 75%. Personal pronouns can also be used as objects to refer to the antecedents which are objects of the previous sentence.

1) Intra-sentential Anaphora Intra-sentential anaphora has the two co-referring expressions in the same sentence [2]. The first phrase in co-reference is called the antecedent and the second one is anaphor. Intra-sentential anaphora resolution relies on syntactic, rather than discourse cues [3].
Ex: 2

a) When jack arrived at the party, he was drunk

b) *When jack arrived at the party she was drunk.*

In the above example 2a) is an ill formed sentence for an obvious constraint, that for two noun phrases to co-refer they must agree in gender, number and person. 1a violates this constraint as jack is female and a pronoun 'he' which is masculine is used to co-refer jack. So the correct usage here is using a pronoun 'she' as in example 2b)

2) Inter-sentential Anaphora: Co reference can occur between two different sentences. If a pronoun is used to refer a noun in the previous sentence it is called inter-sentential anaphoric reference [4]. Pronouns are used to replace nouns. Pronouns have all the features that a noun has. Pronouns carry the information called gender, number and person. They are chosen to refer a noun based on the GNP features of a noun they are referring. Personal pronouns in Telugu corresponding to English [5] [9] are shown in table 1.

TABLE I. PERSONAL PRONOUN OF TELUGU AND ENGLISH

Person	Singular	Plural
1	(I) nEnu (M/F/N)	(We) mEmu (M/F/N)
2	(You) nIvu (M/F/N)	(You) mIru (M/F/N)
3	(He) ataDu (M) (She) Ame (F) (It) adi (N)	(They) (vAru) (They) (avi)

III. RESOLVING THE ANAPHORA

Anaphoric resolution is of crucial importance in order to translate anaphoric expressions correctly into target language. Resolution refers to the process of identifying the antecedent of an anaphor [3]. If there are more than one noun in a sentence to which an anaphor can refer then ambiguity arises in resolving the antecedent of an anaphor. Understanding the sentence and translating them correctly requires world knowledge. Contextual understanding is required to understand and translate such sentences. Humans use more refined and flexible inference making and problem solving capabilities for interpreting these texts. If we can imbibe these processing capabilities to a machine, the accuracy of the system will be as good as a human translator.

Ex: 3

SL: Radha and Ravi are good friends of Raju. She is a very naughty girl.

TL: radha ravi, raju ki ma.nchi snEhitulu. Ame chAla allari pilla.

SL: Raju is playing guitar. It is an electronic device.

TL: Raju guitar vayinchutU unnadu. adi oka electronic parikaram

In the above two sentences the antecedents of anaphors are easily identified. In the first example ‘she’ refers to Radha. In the second example ‘it’ refers to guitar. Translation is done without any ambiguity as these anaphors have only one interpretation in Target language. In some cases there will be more than one antecedent to which an anaphor can refer.

Ex: 4

SL: Radha bought bananas and bangles. They were very sweet.

TL: rAdha araTi paLLu gAjulu techchinadi. avi chAla tiyyaga u.mdinavi

In the above example ‘they’ can refer to either bananas or bangles. GNP features of both the nouns match with GNP features of ‘they’. By applying world and contextual knowledge we can understand that here ‘they’ refers to bananas as bangles have no taste. Translation does not incur ambiguity as both bangles and bananas are having neuter gender, either of them being the antecedent the anaphor ‘they’ will be translated as ‘avi’.

Ex: 5

SL: Radha bought bananas and bangles. They were yellow in color.

TL: rAdha araTi paLLu gAjulu techchinadi. avi pachcha ra.mgu lO u.mdinavi

In the above example ‘they’ can refer to either bananas or bangles. GNP features of both the nouns match with GNP features of ‘they’. By applying world and contextual knowledge we cannot understand whether ‘they’ refers to bananas as bangles as both of them can be yellow in color. Here either the author wants to express that both bananas and bangles are yellow in color or he should explicitly use the noun instead of anaphor to avoid the ambiguity.

Ex: 6

SL: Radha came home with her friends and with some fruits to eat. They look quite tired.

TL: rAdha tana snehitulu mariyu tinuTaku konni paLLu thO vaccinadhi. vAru chAla alisipoyi kanpadutunnaru.

In the above example ‘they’ can refer to either friends or fruits. Number and person features of both friends and fruits are the same but gender feature differ. The gender of friends CAN be male or female and fruits is neuter. ‘They’ can refer either friends or fruits. If ‘they’ refers to a neuter gender noun, then it will be translated as ‘avi’ else it will be translated as ‘vAru/vALLu’ and accordingly the verb suffix will change. By applying world and contextual knowledge its understood that human beings tire but fruits don’t. Accordingly ‘they’ will be translated as ‘vAru’ to refer friends.

Ex: 7:

SL: Radha came home with her friends and some fruits to eat. They were very fresh.

TL: rAdha tana snehitulu mariyu tinuTaku konni paLLu thO vaccinadhi. vAru chala tajaga vu.mdinaru

TL: rAdha tana snehitulu mariyu tinuTaku konni paLLu thO vaccinadhi. avi chala tajaga vu.mdinavi

Taking the same example with slight modification introduces ambiguity. In the example ‘they’ can refer to either friends or fruits as the adjective ‘fresh’ can be used for either of them. By applying world and contextual knowledge Its difficult to tell whether ‘they’ refers to fruits or friends.

IV. TRANSLATION OF ANAPHORS

Translation of anaphors involves three major steps. First step is identification of the antecedent of the anaphor by matching the features of the anaphor with the nouns of the previous in the nominative form. Second step is identification of anaphor of the target language. While translating anaphor from SL to TL the features of anaphors of SL are mapped to the anaphors of target language. If more than one entry is available for the anaphor in the bilingual lexicon then match the GNP features of the antecedent to which the anaphor is referring and anaphor of TL. Third step is verb suffix change according to the subject verb agreement rules of the target language.

- *Verb dependency on Anaphors*

English verbs are not strongly inflected. The only inflected forms are third person singular simple present in –s, a simple past form, a past participle form, a present participle and gerund form in -ing. Most verbs inflect in a simple regular fashion. There are some irregular verbs with irregular past and past particle forms [1]. If pronoun is the subject then the auxiliary verb should agree with the number and person features of the subject.

Telugu verbs are formed by combining roots with other grammatical information. Simple verbs in their finite forms are inflected for tense followed by GNP endings or states. In order to indicate aspect and modality of verbs various auxiliaries are employed

The structure of the verb will be like *Verb stem+ Tense Suffix+ GNP Suffix.* When a pronoun is the subject of a sentence, the verbs agrees in person, number, and when using third person agrees with gender also [7] [8].

The verb inflections should agree with gender and number features of the subject, noun. Though Telugu nouns have three genders and two numbers the verb suffixes change in a different way.

In singular number, feminine and neuter nouns have the same verb suffixes but masculine nouns have different verb suffixes. In plural numbers masculine and feminine nouns have same GNP endings, but for neuter nouns they differ. The suffixes for the verb 'go' are shown in the table 2.

TABLE II. SUFFIXES OF VERB 'GO' FOR DIFFERENT GNP FEATURES

Person	**Singular**		**Plural**	
	Pronoun	**Verb (go/goes)**	**Pronoun**	**Verb (go)**
1	I (nEnu) (M/F/N)	veLLanu	We (mEmu) (M/F/N)	veLLamu
2	You (nIvu) (M/F/N)	veLLavu	You (mIru) (M/F/N)	veLLaru
3	He (ataDu)(M) She (Ame) (F) It (adi) (N)	veLLaDu veLLi.mdi veLLi.mdi	They (vAru) They (avi)	veLLaru veLLayi

TABLE III. VERB PATTERNS OF ENGLISH AND TELUGU

English Pattern	**Telugu Pattern**	**Example Englsh**	**Telugu Translation**
		Single word verbs	
VBD	VBD	He goes	ataDu veLLenu
VBZ	VBZ	We see	mEmu chUsamu
VBP	VBP	I left	nEnu veLLitini
		Two word verb Phrases	
MD+VB	VB+MD	I will stay	nEnu u.mDa galanu
have/has/had+VBN	VBN+ have/has/had	I have gone. She has gone. We had gone	nEnu veLLi unnanu Ame veLLi unnadi mEmu veLLi u.mDagalamu
am +VBG	VBG + am	I am going	nEnu veLLuchU unnanu
is/are +VBG	VBG + is/are	She is going They are going	Ame veLLuchU unnadi vAru veLLuchU unnaru
was/were +VBG	VBG+ was/were	He was going. They were going	ataDu veLLuchU u.mDinADu vAru veLLuchU u.mDiri
am+ VBN	VBN + am	I am done	nEnu chEsinAnu
is/are +VBN	VBN+ is/are	He is released It is taken They are forgiven	ataDu viDudala chEyabaDi u.mnnaDu. Adi tIsukObaDi unnadi vAru kshami.mcha baDi unnAru
was/were +VBN	VBN + was/were	She was forgiven They were forgiven	Ame kshami.mcha baDi u.mDinadi vAru kshami.mchabaDi u.mDiri
		Three word Verb Phrases	
MD+have+VBN	VBN + have + MD	I could have danced	nEnu Adi u.mda galanu
MD+be+VBG	VBG+ be+ MD	She should be arriving	Ame vachuchU u.mDa valenu
MD+be+VBN	VBN+ be + MD	He must be stopped	ataDu Agi u.mDa valenu
have+been+VBG	VBG + been + have	We have been travelling	mEmu prayANamu chEyuchU u.mDi unnamu
has+been+VBG	VBG + been + has	She has been travelling	Ame prayANamu chEyuchU u.mDi unnadi
had+been+VBG	VBG + been + had	It had been raining	ikkaDa varshi.mchuchU u.mDi u.mDagaladu
have+been+VBN	VBN+ been + have	I have been waited	nEnu nirIkshistU u.mDi unnanu
has+been+VBN	VBN+ been + has	She has been tortured	Ame vEdhi.mchabaDi u.mDi unnadi
had+been+VBN	VBN+ been + had	He had been tortured	ataDu vEdhi.mchabaDi u.mDi u.mDagalaDu

English Pattern	Telugu Pattern	Example Englsh	Telugu Translation
am+being+VBG	VBN+ being+ am	I am being groomed	nEnu lali.mchabaDi u.nTU unnanu
is/are+being+VBG	VBN+ being+ is/are	It is being discussed	adi tarki.mchabaDi u.nTu unnadi
was/were+being+VBG	VBN+being+was/were	They were being interrogated	vAru prasni.mchabaDi unTu u.mDiri
Four word verb phrases			
MD+have+been+VBG	VBG+ been+have+MD	It should have been raining	ikkDa varshi.mchuchU u.mDi u.mda valenu
MD+have+been+VBN	VBN+been+have+MD	It should have been rained	ikkaDa varshi.mchabaDi u.mDi u.mDa valenu
MD+be+being+VBN	VBN+being+be+MD	It may be being discussed.	adi tarki.mchabaDi unTu u.mDa galadu

Basic verb phrase patterns in English and their corresponding Telugu translations are shown in below table. From the table below it can be noticed that any verb phrase in Telugu will end with VBD/ VBZ/ VBP/ MD/ have/ has/ had/ am/ is/ are/ was/ were. Depending on the GNP features of the anaphor the last word of a verb phrase should change its suffix [8] [9].

V. Anaphora Resolution Algorithm

The algorithm identifies noun phrase (antecedents) of personal pronouns in English. This work is mainly concentrated on identifying inter-sentential antecedents and is applied to syntactic analysis. Certain constraints and heuristic rules are applied to get a possible solution for an anaphor. Constraints are the GNP agreements of anaphor and the antecedent [3].

Algorithm

```
Step 1: Get the anaphor from a sentence, i.e a word with POS
tag as PP. let it be P1
Step 2: Extract the GNP features of P1 from lexical DB
Step 3: Search for words with NN, NNS, NNP, and NNPS as
their POS tags. Let them be N1, N2...
Step 4: Extract the GNP features of N1, N2...
Step 5: If NP features of P1=3S then
            direct translation
            else if NP features of P1=3P then
           Match the GNP features of P1with N1, N2..
                   If exactly one match found, N1
                           then go to Translation module
                   else if more than one match
                           Apply world and contextual
knowledge to disambiguate.
Step 5: Translation module:
           Get the TL anaphor corresponding to SL anaphor
                   from the bilingual lexicon. Let t      hem be
                   TP1, TP2…
           Match gender feature of TP1, TP2… with N1
                           If Exact match found, TP1
                             Successful translation
                           else
                           no match found for the anaphor.
Step 6: Change the verb suffix according to the GNP features
of TP1.
```

- ***Explanation of Algorithm with examples***

Ex: 8

SL: Students came to the zoo. *They are* watching birds

TL: pillalu ja.mtu pradarshana shAla ki vachiri. *vAru* pakshulanu chUchu *chunnaru*

Ex: 9

SL: Monkeys are in the zoo. *They are* doing mischief

TL: kotulu ja.mtu pradarshana shAla lo vunnavi. *avi* allari cheyu *chunnavi*

In example 1 'they' refers to 'students'. The GNP features of students being (M/F, P, 3), 'they' is translated as 'vAru' and accordingly 'are' is translated as 'chunnaru'. In example 2 'they' refers to monkeys. The GNP features of monkeys being (N,P, 3) 'they' is translated as 'avi' and accordingly 'are' is translated as 'chunnavi'.

For readability purpose the Telugu script in example sentences is transliterated into Roman English using the schema given in Appendix 1.

VI. Conclusion

The algorithm was implemented to translate anaphoric expressions. Personal pronouns whose antecedents precede them were translated successfully.

This work can be extended to deal with cataphoric expressions. This work can be extended to deal with anaphoric expressions having indefinite and reflexive pronouns.

Appendix – 1

Schema for transliterating Telugu as English:
Vowels : అ (a) ఆ (A,aa) ఇ (i) ఈ (I,ii,ee) ఉ (u) ఊ (U,oo) ఋ (RRi) ౠ (RRI) ఎ (e) ఏ (E) ఐ (ai) ఒ (o) ఓ (O) ఔ (au,ou) ; Mathras : ◌ా (A,aa) ◌ి (i) ◌ీ (I,ii,ee) ◌ు (u) ◌ూ (U,oo) ◌ృ (RRi) ◌ౄ (RRI) ◌ె (e) ◌ే (E) ◌ై (ai) ◌ొ (o) ◌ో (O) ◌ౌ (au,ou) ; Anusvara, Visarga and Bindus : ◌ం (.n,.m) ◌ః (H)

Consonants: క (k,q) ఖ (kh,K) గ (g) ఘ (gh,G) ఙ (~N) చ (ch) ఛ (chh,Ch,CH) జ (j) ఝ (jh,Jh,JH) ఞ (~n) ట (T) ఠ (Th,TH) డ (D) ఢ (Dh) ణ (N) త (t) థ (th) ద (d) ధ (dh) న (n) ప (p) ఫ (ph,f) బ (b) భ (bh) మ (m) య (y) ర (r) ల (l) ళ (L) వ (v,w) శ (sh,S) ష (Sh,SH) స (s) హ (h) ;

Extended Consonants: జ఼ (J) క్ష (x) జ్ఞ (GY) ఱ (R)

REFERENCES

[1] Halliday, M.A.K and Hasan, R.1976, Cohesion in English. London: Longman

[2] Webber, B and Reiter, R, Anaphora and Locial form: On Formal Meaning Representations of Natural Language In Proceedings of the fifth IJCAI, Pages 121-131. Cambridge, MA 1977

[3] Shalom Lappin and Herbert J. Leass An Algorithm for Pronominal Anaphora Resolution Journal of Computational Linguistics, Vol. 20, Number 4, 1994

[4] Sidner, C.L, Focusing for Interpretation of Pronouns. Journal of Computational Linguistics 7: 217-231, 1981

[5] English Verbs, Wikipedia, available at http://en.wikipedia.org/wiki/English_verbs

[6] Dr. Divakarla Venkatavadhani, Telugu in Thirty Days, Dakshina Bharat Press, 1976

[7] Albert Henry Arden, A Progressive Grammar of the Telugu Language With Copious Examples And Exercises. India: S.P.C.K Press, 1905

[8] Krishnamurti, B., A Grammar of Modern Telugu. 1985, Delhi; New York: Oxford University Press.

[9] Brown, C.P, The Grammar of the Telugu Language.1991, New Delhi: Laurier Books Ltd.

Learning by Modeling (LbM): Understanding Complex Systems by Articulating Structures, Behaviors, and Functions

Kamel Hashem
Department of Learning Science, School of Educational Sciences
Al-Quds University
Jerusalem, Palestine

David Mioduser
Department of Education in Math Science and Technology, School of Education
Tel-Aviv University
Tel-Aviv, Israel

Abstract—**Understanding the behavior of complex systems has become a focal issue for scientists in a wide range of disciplines. Making sense of a complex system should require that a student construct a network of concepts and principles about the learning complex phenomena. This paper describes part of a project about Learning-by-Modeling (LbM). Many features of complex systems make it difficult for students to develop deep understanding. Previous research indicates that involvement with modeling scientific phenomena and complex systems can play a powerful role in science learning. Some researchers argue with this view indicating that models and modeling do not contribute to understanding complexity concepts, since these increases the cognitive load on students. In this study we investigated the effect of different modes of involvement in exploring scientific phenomena using computer simulation tools, on students' mental model from the perspective of structure, behaviour and function. Quantitative and qualitative methods are used to report about 121 freshmen students that engaged in participatory simulations about complex phenomena, showing emergent, self-organized and decentralized patterns. Results show that LbM plays a major role in students' concept formation about complexity concepts.**

Keywords—learning by modeling; simulation; complexity; mental models; educational technology

I. Introduction

Students' approach to complex dynamic systems and their mental models utilized to construct knowledge, play a powerful role in what students learn. The idea of complexity is increasingly becoming an integral part in learning natural and social sciences, where learning is understood to be more like practice of science [1]. Inquiry-based science, developing skills for systems thinking and adopting collaborative learning in science classes are all examples of that focus.

Students' perception when learning about complex systems is greatly aided by interactive simulations and models. Research indicates that learning through observation do not necessarily lead to strong intuitions or deep understanding of systems [2]. For example people observed bird flocks for thousands of years before anyone suggested that flocks are leader-less, and people participate in traffic jams without much understanding of what cause the jams, such phenomena may be regarded as complex systems. Observation and participation are not enough; people need a richer sense of involvement with systems in order to understand them [3], [4], [5], [6], [7], [8], [9]. Modeling can provide students with the power to understand and explore systems that were previously difficult to trace and predict their behavior, new techniques that help to learn important concepts on complex systems, to generate relevant questions, theories and hypothesis about phenomena, and to build and run models related to their theories [10], [11], [12], [13], [14].

Emergent complex phenomena are considered to be difficult to understand [15], [16]. Despite the utilization of new learning approaches with models, students experience difficulties in learning concepts relevant to understanding complex systems currently taught in existing science courses – student thinking may be counter-intuitive or might conflict with the scientific models, and the learning ideas concerning emergence or stochastic processes are difficult because of difference with teleological beliefs, where students tend to think of systems having centralized control [14], [15], [16], [17]. Hmelo-Silver and Pfeffer (2004) argue that the characteristics of complex systems make them difficult to understand, since they are comprised of multiple levels of organization that often depend on local interactions (the causes and effects are not obviously related); also it requires that students should construct a network of concepts and principles about the phenomena with complexity and their interrelationships [14], [16], [18].

Some researchers argue that modeling did not contribute a lot in understanding complexity since it increases the cognitive load on students (see [19]). This study focused on the effect of different modes of involvement in exploring scientific phenomena using computer modeling tools, on students' mental model from the perspective of the system structure, system behavior and function. It is part of a more comprehensive study pursuing the goals: (1) to study the role of modeling in the learning process of complexity and complex systems in the natural and artificial worlds; (2) to examine the contribution of different modes of involvement in the modeling process (e.g., observation and explanation, intervention and manipulation, programming and development) to the students' understanding of complexity; (3) to examine the effect of the level of complexity and properties (e.g., emergence, self-organized …) of the systems being manipulated on the student's learning; and (4) to study the evolution in time of the

TABLE I. CATEGORIZATION OF CSMM ACROSS SBF CONCEPTS

Mental Model	Questions
Structure	Describe what you see in detail (number of agents, how do agents behave before they are part of the system, system environment)?
Function	Who/what initiates the formation of the system? Are there feedback loops within the system? Do they amplify or control the outcome? How do agents behave before they are part of the system? Is the same outcome will be achieved each time the system form? How would the system respond to environmental change, explain why?
Behaviour	Is there movement of the agents within the system? How would you design such system/explain its behavior? Is there a difference between agents and system? What draws the system together?

students' mental models of complexity as a function of the different variables (e.g., modes of involvement; level of complexity) of the system under study.

II. METHOD

A. Subjects

Participants are 121 undergraduate students (ages ranging from 18 to 20 years old) from the science department at Al-Quds University in Jerusalem, divided into four groups by the kind of involvement in working with models: observation, exploration, manipulation, and model-development modes. All students attended a two hours introduction lesson to the NetLogo environment. The students were selected based on their scientific background, all have done the tawjehi exam as required by the ministry of education for the scientific track, and they are all studying first year compulsory science courses in the faculty of science.

B. Research instruments

(a) The learning environment comprising two components: (1) NetLogo, a specialized program developed at Northwestern University for agent-based modeling and for learning and understanding complex systems, "NetLogo is a multi-agent programming language and modeling environment for simulating natural and social phenomena. It is particularly well suited for modeling complex systems evolving over time. Modelers can give instructions to hundreds or thousands of independent "agents" all operating concurrently. This makes it possible to explore connections between micro-level behaviors of individuals and macro-level patterns that emerge from their interactions, it enables users to open simulations and "play" with them, exploring their behavior under various conditions. NetLogo is also an authoring environment that is simple

TABLE II. CATEGORIZATION OF CONCEPTS RELATED COMPLEX SYSTEMS MENTAL MODELS (CSMM)

Parameters	Clockwork component coding (reductive)	Complexity component coding (non-reductive)
System control 1. Who/what initiates the formation of the system?	**Centralized** Order/control come from outside.	**Decentralized** Agents' actions are independent of each other; they operate under the same rules.
Action effects 1. Are there feedback loops within the system? 2. Do they amplify or control the outcome?	**Linear** One thing leads to another, direct link between cause and effect.	**Non-linear** Positive feedback can exhibit exponential results. Effects are not straightforward functions of causes
Agents' action 1. How do agents behave before they are part of the system?	**Predictable** Agents' actions are predictable; there is no mention of randomness or chance in their action.	**Random** 1. Agents appear to act in random independent fashion. 2. Randomness allows for variability and variety within the system.
Underlying causes 1. Is the same outcome will be achieved each time the system form? 2. How would the system respond to environmental change, explain why?	**End point is predictable (teleologic)**	**Probabilistic causes (stochastic)** 1. The system organizes itself based on agents interactions, the resulting structure is never certain. 2. The system maintains its coherence/structure.

enough to enable students and researchers to create their own models, even if they are not professional programmers", in a Netlogo environment, students come to understand these concepts and laws through a process of exploration and inquiry by investigating and controlling the behavior of thousands of graphical "agents". By interactively exploring the relationship between the agents' rules of behavior and the patterns that emerge as a result of these rules, students are able to "debug" misconceptions that are generated by confusing their understandings of micro and macro level interactions. Typically, in curricula using multi-agent modeling, students begin by exploring the behavior of pre-built simulations designed to focus on some target concepts. They make predictions about the behavior of the model under varying model parameters then test their predictions by exploring model outcomes as they change 'sliders' in a simple graphical user interface (see figure below). The core of every NetLogo model is the *interface window*. Typically, the interface contains a graphics window, a plotting window and several variables in the form of sliders and buttons that the student can manipulate. It is here that students can observe directly the interaction between the micro- and macro-levels". [20: 1-2].

and (2) tasks and activities in which students run NetLogo models and are requested to perform tasks with the models, and

(b) Data collection tools included: (1) pre-test comprising general background and demographic information (e.g., major area, gender) and four questions dealing with complexity concepts such as emergence, self-organization and decentralization; (2) structured observation and data forms; (3) mental model worksheet focusing on students' complex-system-mental-model (CSMM), completed by the end of each activity.

C. Procedure

The study was carried in four stages: (a) Pre-test, (b) Treatment in four different modes. *Observation group*: in 2 sessions of 90-minutes students were introduced to two models (two levels of complexity), requested to observe agents interactions, and to complete the CSMM worksheet. *Exploration group*: in 2 sessions of 90-minutes students were introduced to two models (two levels of complexity), were given an initial set of conditions for the system followed by a final set of conditions in which one or two parameters were changed (e.g., change in a variable-slider or a switch) while the others remained constant. After each model students were interviewed and requested to complete the worksheet. *Manipulation group*: in 2 sessions of 90-minutes students were introduced to two models (two levels of complexity) using NetLogo. They were asked about how the system would change if the system variables were altered, and even allowed to use NetLogo commands. The students then manipulated the system variables according to the interviewer's questions, explained their observations of the system's behavior and compared these with their initial predictions. After each model students completed the worksheet. *Development and Design group*: students were introduced to the Netlogo programming environment (48 hours), in order to have the ability to construct the learning models. After each model students were asked to complete the worksheet, (c) Interview after treatment: students were interviewed for their CSMM, all responses were audio taped and (d) Post-test: (same as pre-test).

D. Scoring

First coding scheme was based on the distinction between a system's structure, behavior, and function (SBF) (see Table I). Structure refers to elements of a system and their configuration (e.g., agents, environment, and interaction between them); behavior refers to how systems achieve their purpose through the interactions or of its agents and Function refers to the purpose of the agents in a given system. The second coding scheme was based on the categorization defined by Jacobson (2001), shown in Table II.

Students' answers were coded as non-reductive if these referred to a complexity-related matter (i.e. the whole is greater than the parts). Otherwise, if there was evidence of a stepwise approach to the explanation, the answer was coded as reductive (i.e. agents act in isolation). Jacobson (2001) refer to the reductive way of thinking as "deterministic and clockwork order".

III. Results

A. Quantitative analysis

Student's responses were coded in terms of the various types of component beliefs reflected in their answers, for the pre-post test results showed an increase in students' understanding of complexity concepts in all the four groups (observation, exploration, manipulation, and development and design) (see Hashem and Mioduser, 2011).

TABLE III. Students' CSMM (frequencies and percentages) across SBF concepts for the different modes of involvement (** p<0.01)

Modes of Involvement		Structure ** Concept presence		Function **		Behaviour **	
	(N)	No	Yes	Clockwork	Complex	Clockwork	Complex
		Frq (%)	Frq (%)	Frq (%)	Frq (%)	Frq (%)	Frq (%)
Observation	58	32 (18)	142 (81)	162 (55)	126 (43)	118 (67)	40 (22)
Exploration	66	11 (5)	187 (94)	143 (43)	185 (56)	117 (59)	75 (37)
Manipulation	56	8 (4)	160 (95)	103 (36)	173 (61)	91 (54)	76 (45)
Design	62	11 (5)	175 (94)	90 (29)	217 (70)	76 (40)	101 (54)

TABLE IV. : STUDENTS’ MENTAL CONCEPTS ACROSS SBF CONCEPTS VS. MODES OF INVOLVEMENT (* P < 0.05, ** P < 0.01)
TABLE V.

Complexity Level	MM concept	Response	Group			
			Observation	Exploration	Manipulation	Design
			Frq (%)	Frq (%)	Frq (%)	Frq (%)
Complicated	Structure ** - concept presence	No	15 (17.2)	4 (4)	3 (3.6)	6 (6.5)
		Yes	72 (82.5)	95 (96)	81 (96.4)	87 (93.5)
	Function **	Clockwork	79 (55.2)	78 (47.6)	43 (31.6)	45 (29.4)
		Complex	64 (44.8)	86 (52.4)	93 (68.4)	108 (70.6)
	Behavior **	Clockwork	59 (72.8)	63 (67)	44 (53)	42 (47.7)
		Complex	22 (27.2)	31 (33)	39 (47)	46 (52.3)
Complex	Structure ** - concept presence	No	17 (19.5)	7 (7.1)	5 (6)	5 (5.4)
		Yes	70 (80.5)	92 (92.9)	79 (94)	88 (94.6)
	Function **	Clockwork	83 (57.2)	65 (39.6)	60 (42.9)	45 (29.2)
		Complex	62 (42.8)	99 (60.4)	80 (57.1)	109 (70.8)
	Behavior **	Clockwork	59 (76.6)	54 (55.1)	47 (56)	34 (38.2)
		Complex	18 (23.4)	44 (44.9)	37 (44)	55 (61.8)

As expected, students in the development and design group identified more concepts across the structure, function and behavior (SBF) framework than the other groups. A general log-linear analysis was conducted to examine the differences between the groups in their representation on structures, behaviors, and function, showing significant interaction between the modes of involvement and SBF concepts (χ2 (df = 25) = 100.860, p < 0.01).

Table III presents the frequencies and percentages of the students’ responses on the CSMM across the SBF concepts for the four groups (observation, exploration, manipulation and design). A chi-square test was done to check these frequencies for significance; results show a significant relationship between the different modes of involvement and the CSMM across the SBF concepts as follows: (a) Students largely identified a target list of questions regarding structure in the complex system mental model (CSMM) (χ2 (df = 3) = 28.588, p < .01), in examining the observed cell frequencies from Table III, it shows that that the manipulation group got the highest frequency (95%) in identifying the concepts regarding system, followed by the design group and the exploration group (94%) and finally the observation group (81%), (b) Students largely favor to choose the clockwork model when they were asked questions regarding system functioning (χ2 (df = 3) = 47.151, p < .01), in examining the observed cell frequencies from Table III, the observation group showed the highest frequency (55%) in favoring the clockwork model, on the other hand the design group showed high response in choosing the complex model on system functioning (70%) followed by the manipulation group (61%) followed by the exploration group (56%) and finally the observation group (43%), and (c) Students largely favor to choose the clockwork model when they were asked questions regarding system behavior (χ2 (df = 3) = 36.043, p < .01), in examining the observed cell frequencies from Table III, the observation group showed the highest frequency (67%) in favoring the clockwork model followed by the exploration group (59%) followed by the manipulation group (54%) and finally the design group (40%), on the other hand the design group showed the highest response in choosing the complex model on system behavior (54%) followed by the manipulation group (45%) followed by the exploration group (37%) and finally the observation group (22%).

The awareness regarding complex system mental model (CSMM) that was mentioned in Table II, can be seen across the different mental concepts: structure, function and behavior (SBF) that was mentioned in Table I with different complexity levels in Table IV, the different groups showed a significant interaction with the CSMM across the SBF concepts while interacting with models of different complexity levels. A general log-linear analysis was conducted using SPSS software to examine the differences between complexity levels (complicated and complex) and students perception on the CSMM across SBF concepts showing significant relationship (χ2 (df = 25) = 68.769, p < 0.01) for complicated model and (χ2 (df = 25) = 65.517, p < 0.01) for complex model.

Table IV and Figure 1 presents the frequencies and percentages of the students’ responses on the CSMM across SBF concepts with different complexity levels for the four groups (observation, exploration, manipulation and design). A chi-square tests was done to check these frequencies for significance, results show a significant relationship between the different complexity levels and the different complexity concepts as follows:

(1) For the complicated model: (a) Students largely identified a target list of questions regarding *structure* in the model under study (χ2 (df = 3) = 15.204, p < .01), in examining the observed cell frequencies from Table IV, it shows that that the manipulation group got the highest frequency (96.4%) in identifying the model structure, followed

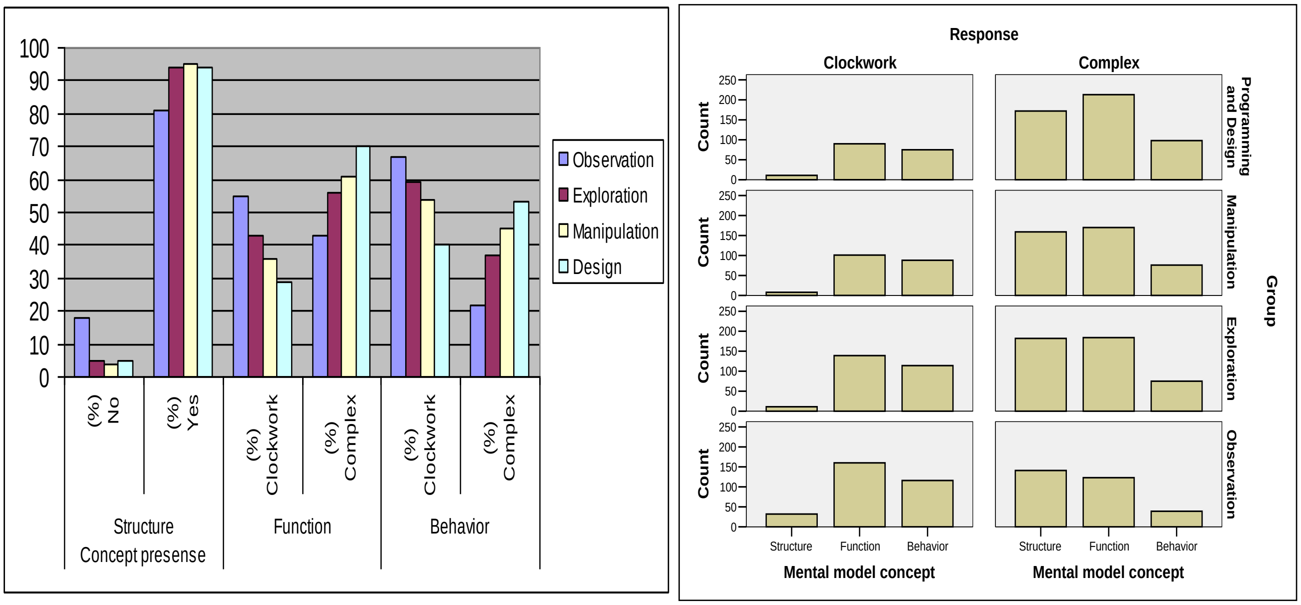


Fig.1. STUDENTS' CSMM ACROSS THE SBF LAYERS FOR THE DIFFERENT LEVELS OF INVOLVEMENT

by the exploration group (96%) followed by the design group (93.5%) and finally the observation group (82.5%), (b) Students largely favor to choose the clockwork model when they were asked questions regarding system *functioning* (χ2 (df = 3) = 28.331, p < .01), in examining the observed cell frequencies from Table IV, the observation group showed the highest frequency (55.2%) in favoring the clockwork model, on the other hand the design group showed high response in choosing the complex model on system *functioning* (70.6%) followed by the manipulation group (68.4%) followed by the exploration group (52.4%) and finally the observation group (44.8%), and (c) Students largely favor to choose the clockwork model when they were asked questions regarding system *behavior* (χ2 (df = 3) = 14.718, p < .01), in examining the observed cell frequencies from Table IV, the observation group showed the highest frequency (72.8%) in favoring the clockwork model followed by the exploration group (67%) followed by the manipulation group (53%) and finally the design group (47.7%), on the other hand the design group showed the highest response in choosing the complex model on system *behavior* (52.3%) followed by the manipulation group (47%) followed by the exploration group (33%) and finally the observation group (27.2%), and (2) For the complex model: (a) Students largely identified a target list of questions regarding *structure* in the model under study (χ2 (df = 3) = 14.120, p < .01), in examining the observed cell frequencies from Table IV, it shows that that the design group got the highest frequency (94.6%) in identifying the model structure, followed by the manipulation group (94%) followed by the exploration group (92.9%) and finally the observation group (80.5%), (b) Students largely favor to choose the clockwork model when they were asked questions regarding system *functioning* (χ2 (df = 3) = 24.577, p < .01), in examining the observed cell frequencies from Table IV, the observation group showed the highest frequency (57.2%) in favoring the clockwork model, on the other hand the design group showed high response in choosing the complex model on system *functioning* (70.8%) followed by the exploration group (60.4%) followed by the manipulation group (57.1%) and finally the observation group (42.8%), and (c) Students largely favor to choose the clockwork model when they were asked questions regarding system *behavior* (χ2 (df = 3) = 24.726, p < .01), in examining the observed cell frequencies from Table IV, the observation group showed the highest frequency (76.6%) in favoring the clockwork model followed by the manipulation group (56%) followed by the exploration group (55.1%) and finally the design group (38.2%), on the other hand the design group showed the highest response in choosing the complex model on system *behavior* (61.8%) followed by the exploration group (44.9%) followed by the manipulation group (44%) and finally the observation group (23.4%).

B. *Qualitative analysis*

An examination of students' responses indicated additional qualitative differences between the different modes of involvement. The programming and design group provided more elaborate responses as well as demonstrating more understanding across the SBF concepts followed by the manipulation group followed by the exploration group and finally the observation group, this was evident in their answers. All the groups have identified the various system structures, but on the behavioral and functional level the programming and design group have discussed in more details. For example, in a description for the traffic jam model, one of the students in the design group said:

....The system consists of a number of cars that are driving in different velocities along the street…. The traffic jam occurs when we have an increase in the number of cars, specially the number of the private ones, and there is a feedback loop since the movement of each car is affected by the car in the front and in back…. The system draws together because of the restricted movement for the cars and all behavior looks semi-organized, where the elements of the system work in consistent to perform the target objective….

In this example the student mentions the system structure consisting of (cars, street) and continues to discuss the function of the cars and how traffic jam occurs because of the number of cars (non-reductive), and the system achieve its purpose through the restricted interactions between the cars and all behavior looks semi-organized (non-reductive). A student from the observation group responded to the same instructions with the following:

..... The system consists of a group of cars and street all the cars are blue; only one is red.... one of the things that might impede the flow of traffic is the traffic lights or an accident.... No traces for feedback since the cars are going at same speed...

This student mentioned numerous structures (cars, street, and agents color), she describes the functionality of the system in a reductive way (no feedback loops, cars are going at same speed) and did not offer additional behavioral information. Integrated students responses were also evident in the interviews where students were asked about the traffic jam formation and how it occurs? For example in a response to this question one of the students in the observation group noted:

S: "First of all you might have an accident in the road or the road is not good to let drivers pass in a regular way…"

I: “ok, suppose we have no accidents and the road is good, is there any chance to a have a traffic jam?”

S: "Yes, let’s say if we have a traffic light…"

In this answer, the student’s response has been coded as ‘reductive or clockwork’ because it referred to a centralized control and deterministic single causality (i.e. the references to ‘accidents, road is not good, and traffic light’).

Once again, students in the design group include more structural, functional and behavioral non-reductive responses in their answers followed by the manipulation group followed by the exploration group and finally the observation group, most of the groups indicated high responses in identification of the various system structures for both complicated and complex systems.

IV. DISCUSSION AND CONCLUSION

Results showed that learning about complicated systems is not easier than learning about complex system with all groups. One of the reasons may relate to the specific model they interacted with.

Some students found it difficult trying to understand the concept of emergence, when they were involved with the complicated model “chemical equilibrium”. In the task, they focused on the actions of molecules at the micro level (i.e., how molecules move, interact…etc.), rather than noticing the collective interactions of all the molecules. This means that they did not focus on the sum of all interactions of the individual molecules across time. They aknowledged that these reach the equilibrium state from observing the statistics regarding number of interacting and yeilding molecules.

When involved with the complicated model, students in the manipulation group showed elaboration and understanding in terms of the approach unveiled (e.g., "clockwork" vs. "complex") for the concepts of emergence and self-organization.

In contrast, when working on the complicated model, the exploration group showed an elaboration and understanding on “the way that systems are governed”. Their description was that the actions of the actors are not directed towards any goal and molecules A and B don’t have to interact or achieve an equilibrium, they simply move around and collide randomly [2].

Learning by modeling provided the base for promoting doing with reflection and for helping students make connections to the world around them [6]. The groups’ differences unveil which complexity level was better understood by the students. We can say that students involvement with systems with complexity level of 'complex' have elaborated and understood more complex systems concepts while interacting with models than the students who were involved with systems with complexity level of 'complicated'.

Ultimately this article reports on a study about the interaction between modes of learning with computer modeling tool and the understanding of complexity concepts, there are many systems concepts that we never directly experience or that violate our intuitions and challenges of our cognitive and metacognitive resources. The implementation of such an instructional approach in the curriculum would have many benefits for learners, such as interdisciplinary learning allowing to see common patterns across traditionally separate fields, new ways of thinking (systems thinking and decentralized thinking), exploration of tools to think with, and construction of models linking between local causes and global behavior.

By introducing this new perspective (LbM) using computer modeling for learning complexity and emergent phenomena, science learning will be more motivational and truthful, more inclusive and accessible to the great majority of students, the use of the SBF framework allows effective reasoning about the structural, behavioral and functional roles within the system under study, in addition, this study's results have clear implications for the design of learning environments that can support learning about complex systems.

REFERENCES

[1] D. Chen, and W. Stroup, “General system theory: Toward a conceptual framework for science and technology education for all,” *Journal of Science Education and Technology*, 2 (3), 447-459, 1993.

[2] M. Resnick, “Beyond the centralized mindset,” *Journal of the Learning Sciences*, 5 (1), 1-22, 1996.

[3] K. J. Gilbert, and J. C. Boulter, (Eds.), “Developing models in science education,” Dordrecht, Holland: Kluwer Academic Publishers, 2000.

[4] D. J. Gobert, and C. B. Buckley, “Introduction to model-based teaching and learning in science education,” *International Journal of Science Education*, 22 (9), 891-894, 2000.

[5] L. Louca, and C. Constantinou, “The use of computer-based microworlds for developing modeling skills in physical science: an example from light,” *International Journal of Science Education*, 2003.

[6] C. Hmelo-Silver, and M. G. Pfeffer, “Comparing expert and novice understanding of a complex system from the perspective of structures, behaviours, and functions,” *Cognitive Science* , 28 (1), 127-138, 2004.

[7] K. Hashem, and I. Arman, “Integration of ICT in mathematical understanding using modeling,” *International Journal of Computerand Information Technology*, 2(2), 330-335, 2013.

[8] M. Resnick, and U. Wilensky, “Diving into Complexity: Developing probabilistic decentralized thinking through role-playing activities,” *Journal of Learning Sciences*, 7 (2), 153-172, 1998.

[9] C. Yehezkel, M. Ben-Ari, and T. Dreyfus, “Computer architecture and mental models,” *ACM* , 101-105, 2005.

[10] P. Blikstein, and U. Wilensky, “Less is more: agent-based simulation as a powerful learning tool in materials science,” *Proceedings of the IV International Joint Conference on AAMAS*. Utrecht, Holland, 2005.

[11] K. Hashem, and D. Mioduser, “The Contribution of Learning by Modeling (LbM) to Students' Understanding of Complexity Concepts,” *International Journal of e-Education, e-Business, e-Management and e-Learning (IJEEEE)* , 1 (2), 151-155, 2011.

[12] S. Levy, and U. Wilensky, “An analysis of student patterns of exploration with NetLogo models embedded in the connected chemistry environment,” *Proceedings of the annual meeting of the American Educational Research Association*. Montreal, CA, 2005.

[13] M. Stieff, and U. Wilensky, “Connected chemistry-incorporating interactive simulations into the chemistry classroom,” *Journal of Science Education and Technology*, 12 (3), 285-302, 2003.

[14] U. Wilensky, and M. Resnick, “Thinking in levels: A dynamic systems approach to making sense of the world,” *Journal of Science Education and Technology* , 8 (1), 3-19, 1999.

[15] M. Jacobson, “Problem solving, cognition, and complex systems: Differences between experts and novices,” *Complexity*, 6 (3), 41-49, 2001.

[16] M. Jacobson, and U. Wilensky, “Complex systems in education: scientific and educational importance and implications for the learning sciences,” *Journal of the Learning Sciences* , 15 (1), 11-34, 2006.

[17] M. T. Chi, “Commonsense conceptions of emergent processes: why some misconceptions are robust?” *The Journal of the Learning Sciences* , 14 (2), 161-199, 2005.

[18] M. Resnick, “Changing the centralized mind. Cambridge,“ MA: MIT press, 1994.

[19] J. Gobert, “Harnessing technology to support on-line model building and peer collaboration,” 2003.

[20] S. Tisue, and U. Wilensky, “NetLogo: A Simple Environment for Modeling Complexity,” *International Conferece on Complex Systems*. Boston, 2004.

A Review of Computation Solutions by Mobile Agents in an Unsafe Environment

Anis Zarrad
Department of Computer Science and Information Systems
Prince Sultan University, Riyadh,
Saudi Arabia

Yassine Daadaa
College of Computer and Information Sciences
Al-Imam Muhammad ibn Saud Islamic University, Riyadh,
Saudi Arabia

***Abstract*—Exploration in an unsafe environment is one of the major problems that can be seen as a basic block for many distributed mobile protocols. In such environment we consider that either the nodes (hosts) or the agents can pose some danger to the network. Two cases are considered. In the first case, the dangerous node is a called black hole, a node where incoming agents are trapped, and the problem is for the agents to locate it. In the second case, the dangerous agent is a virus; an agent moving between nodes infecting them, and the problem is for the "good" agents to capture it, and decontaminate the network.**

In this paper, we present several solutions for a black–hole and network decontamination problems. Then, we analyze their efficiency. Efficiency is evaluated based on the complexity, and the effort is in the minimization of the number of simultaneous decontaminating elements active in the system while performing the decontamination techniques.

Keywords—Distributed algorithm; Mobile Agent; Network Decontamination; Black Hole Search; and Network Exploration

I. Introduction

Today's need to maintain network protection practices and challenges in the universe are interconnected. Virus protection represents a rising importance in network decontamination methods. Faults and viruses often spread in networked environments, where nodes represent hosts and edges represent connections between hosts, by propagating from neighboring sites.

Such a topic is known as exploration in unsafe environment. Exploration consists of having a set of agents collaboratively traverse an unknown network to collect relevant information. Network exploration has been extensively studied over the past fifty years due to its various applications in different areas such as engineering, computer science, and applied mathematics. Two major problems are discussed in this work; black-hole and network decontamination.

In network contamination, a node might behave incorrectly, and it could affect its neighbor to become contaminated as well, thus propagating faulty computations. The propagation patterns of faults can follow different dynamics, depending on the behavior of the affected site.

In this work we begin by giving a general overview about network exploration in unsafe environment and its applications. The rest of the paper is organized as follows. First, we summarize the backgrounds and related works, second we review some of the major solutions to date and their classification, and finally in section four we offer a conclusion.

II. Backgrounds And Related Works

The main focus in this work is computation solutions by mobile agents in an unsafe environment. Also, we believe there is a need to highlight the most influential works related to mobile agent systems that occurred in the past in order to give credit to founder researches. The problem of exploration is well known, where agents need to collaborate in order to explore an unknown environment. For example, tasks in environments that is not suitable for human operation, navigating a robot through a terrain containing obstacles, and finding a path through a maze. In recent years, application such as searching for data stored at unknown nodes in a computer network using mobile software agents, and obtaining maps of existing networks (e.g., computer networks, sewage systems, unexplored caves) have been studied. Map construction is the related problem of exploring the network to return an exact map of its topology.

Previous work on exploration of labeled graphs has emphasized minimizing the cost of exploration in terms of the total number of edge traversals (moves), and the amount of memory used by the agent [1, 2, 12, 13, 38]. In [2], Awerbuch et al. studied how a mobile robot can learn an unknown environment in a piecemeal manner. The robot's goal is to learn a complete map of its environment, while satisfying the constraint that it must return every so often to its starting position (e.g., for refuelling).

Exploration of anonymous graphs is impossible if marking of the nodes is not allowed in some way. An exception is when the graph is acyclic, meaning the graph is a tree [14, 30]. Different models for marking the nodes have been used to solve the exploration problem. Pebbles which can be dropped and removed from a node was proposed by Bender et al. in [5], where it was shown that one pebble is enough to explore the graph if the robot knows an upper bound on the size of the graph, and (log log n) pebbles are necessary and sufficient otherwise. Among the various possible techniques of decontamination, two types have been identified in the literature [24] internal and external decontamination.

In internal decontamination a site can decontaminate itself (i.e., it can activate an antiviral software) when a certain condition of the neighborhood is verified. A clean site gets re-contaminated when some other condition of the neighboring states is verified. This approach has been followed in [39], where a node becomes clean when the majority of its neighbors are clean, and a clean node becomes contaminated if any of its neighbors are contaminated. A similar approach has been taken in [36, 37], where a decontaminated node is immune to recontamination if the majority of its neighbors are clean. Internal decontamination has been specifically studied in the context of fault- tolerance to describe the mechanisms of the spread of faults and of auto-correction (see [40] for a survey). In all these studies the main objective is typically to determine the minimum size of a set of faulty nodes, which completely disrupts the system under given contamination/decontamination dynamics or, equivalently, the minimum size of a set of decontaminating nodes that can decontaminate the whole network under the same circumstances.

On the other hand, mobile agents moving in the network can perform external decontamination. There is an extensive literature on external decontamination either in specific topologies (see [3, 25]), under various assumptions on the capabilities of the agents, or in arbitrary topologies (see [7]). Typically, agents have memory, distinct identifiers, and the ability to communicate with other agents when they meet or exchange information writing on whiteboards (storage area located at the nodes). In all models investigated, agents can move from node to node (usually asynchronously and independently) decontaminating the sites they pass through. A clean site becomes contaminated if at least one of its neighbors is contaminated. External decontamination has been studied in the context of intruder capture to design algorithms to neutralize a virus in a network, or in graph search (e.g., [3, 31, 25, 6, 4]). The main goal of decontamination in these settings is usually to design a strategy that employs the minimum possible size of the team of cleaning agents.

III. Black Hole Search – A Classification Review

In the last decade, there has been a great deal of work done on finding faults in networks using mobile entities. Most of the existing work deals with the black hole search problem for a single black hole [9, 10, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 32, 33, 34]. Some work has been done for multiple black holes in [8, 35], although both of these papers use the synchronous model, as do some of the single black hole papers [9, 10, 33, 34]. Existing solutions can be classified based on the network model: synchronous and asynchronous.

A. Synchronous Model

In [33], Klasing et al. consider the problem of designing a black hole scheme under the scenario of synchronous networks. Authors investigate the case when there may be at most one black hole in the network. The search is performed by exactly two agents, which start from the same node home-base and can communicate only when they are in the same node. At least one agent must report the information back to the home-base on the exact location of the black hole or whether one exists. In [9], Czyzowicz et al. studied the black hole problem in a (partially) synchronous network, assuming an upper bound on the time of any edge traversal by an agent. In this work, the minimum number of agents capable to identify a black hole is two for a given graph and a given starting node.

Czyzowicz et al. [9] looked to the fastest possible black hole search by two agents, under the general scenario in which some subsets of nodes are safe and the black hole can be located in one of the remaining nodes. Authors show that the problem of finding the fastest possible black hole search scheme by two agents is NP-hard, and they give a polynomial approximation to solve it.

In [10], Czyzowicz looked at the same problem in trees, and gave optimal black hole search algorithms for two extreme classes of trees: the class of lines and the class of trees in which any internal node (including the root which is the starting node) has at least two children. In [34], Klasing et al. consider the same problem assuming that the map of the network is given. Their objective is minimizing the number of agents that fall into the black hole and the time taken by the surviving agents to locate the black hole. The proposed algorithm explores the network via a spanning tree. In [8], Cooper et al. were the first to consider the general case of multiple black holes using k agents starting from the same node. The agents move through the network in synchronous steps and can communicate only when they meet in a node. In [35], Kosowski et al. look at a multiple black hole search: assuming a directed graph. The robots are associated with unique identifiers, they know the number of nodes in the graph (or at least an upper bound), and they know the number of edges leading to the black holes. Each node is associated with a whiteboard, separately considering the synchronous and the asynchronous cases.

B. Synchronous Model

The asynchronous scenario was studied under different agent models (i.e., network knowledge, tokens, and whiteboard). In [19], Dobrev *et al*. provided solutions to the black hole search problem in anonymous rings using whiteboards for two settings; when the anonymous agents are co-located, and when they are dispersed. They proved that two such agents are necessary and sufficient to locate the black hole. In [18], Dobrev *et al*. provide a characterization of the impact that factors such as a priori network knowledge and consistency of the local port labeling have on the complexity of the black hole location problem. In [18] authors consider both the case of topological ignorance in systems where there is sense of direction and the case of complete topological knowledge of the network. Authors show that, in both cases, two agents suffice.

In [15, 20, 21], Dobrev *et al.* investigate the black hole search problem for two agents with a map and whiteboard searching for a single black hole. In [20], authors present a search protocol that improves the bounds from the worst case lower bound of $\Omega(\mathrm{n}\log n)$ agents moves to $O(n + d\log d)$ agents moves, where d is the diameter of the network. The result allows for $\Theta(n)$ moves for a large class of possibly unstructured networks with low diameter. In [15], Dobrev *et al.* show that with a map of the network, a team of two agents suffices, and the number of moves is in the worst case $\mathrm{O}(\mathrm{n}\log n)$. They also present a general strategy that allows two agents to locate the black hole with $\mathrm{O}(\mathrm{n})$ moves in common interconnection networks such as hypercubes, cube connected cycles, star graphs, wrapped butterflies, and chordal rings, as well as in multidimensional meshes and tori of restricted diameter. These results hold even if the networks are anonymous. In [19], authors use a technique based on pre-calculating the open vertex cover of cycles of a graph that allows them to solve the problem in $\Theta(n)$ for a large class of networks.

In contrast, in [32], Glaus studied the black hole search problem without the knowledge of the incoming link, and has shown that this modification has effects on the size of the solution. Glaus [32] provided the lower bound on the number of agents that are necessary to locate the black hole; any correct algorithm solving the black hole search problem without the knowledge of the incoming link needs at least $\frac{\Delta^2+\Delta+2}{2}+1$ agents. The algorithm uses the optimal number of agents in the worst case, however, the cost of the algorithm and bounds on the optimal cost of the solution were not shown in this paper.

In [22], Dobrev et al. consider the token model, where each agent has a bounded number of tokens available that can be carried, placed on, or removed from a node. All tokens are identical (i.e., indistinguishable), and no other form of communication or coordination is available to the agents. Authors first prove that a team of two agents is sufficient to locate the black hole infinite time even in this weaker coordination model. Furthermore, Dobrev et al. [22] prove that this can be accomplished using only $O(n \log n)$ moves in total, which is optimal, the same as with whiteboards. Finally, authors show that to achieve this result the agents need to use only $O(1)$ tokens each.

The previous strategy is generalized in [16], where Dobrev *et al.* look to the case of unknown graph. Authors present an algorithm that works in the token model and solves the black hole search problem with the minimal number of agents and with a polynomial number of moves. Dobrev *et al.* [16] algorithm works even if the agents are asynchronous, and if both the agents and the nodes are anonymous. More precisely, authors consider an unknown, arbitrary, anonymous network and a team of exploring agents starting their identical algorithm from the same node (home-based). The agents are anonymous, and they move from node to neighboring node asynchronously. Each agent has an indistinguishable token (or pebble) available that can be placed on, or removed from a node. The token can be placed on a node, either in the center or on an incident link. In the proposed algorithm, two tokens are never placed in the same location (node center or port), nor does an agent ever carry more than one token. Using only this tool for marking nodes and communicating information, authors show that with $(\Delta + 1)$ agents (where Δ is the maximal degree of the graph), the exploration can be successfully completed. The proposed algorithm allows at least one agent to survive and, within a finite time, the surviving agents will know the location of the black hole with the allowed level of accuracy. The number of moves performed by the agents when executing the proposed protocol is shown to be polynomial, and the proposed algorithm is rather complex.

In [23], Dobrev et al. show not only that a black hole can be located in a ring using tokens with scattered agents, but also that the problem is solvable even if the ring is unoriented. First authors prove that the black hole search problem can be solved using only three scattered agents. Then, Dobrev et al. [23] show that, with k (k > 4) scattered agents, the black hole can be located in O(kn+n log n) moves. Moreover, when k(k > 4) is a constant number, the move cost can be reduced to O(nlogn), which is optimal. These results hold even if both agents and nodes are anonymous.

In [17], Dobrev et al. consider k anonymous, asynchronous mobile agents in an anonymous ring with a black hole. The agents are aware of the existence, but not of the location of such a danger, and the network are totally asynchronous. In this setting it was observed that in order to solve the problem, the network must be 2-connected. A black hole search is not feasible in trees, because in asynchronous networks it is impossible to distinguish a black hole from an incident slow link. The only way to locate a black hole is to visit all other nodes and learn that they are safe. In particular, it is impossible to answer the question of whether a black hole actually exists in the network, hence authors worked under the assumption that there is exactly one black hole and the task was to locate it. In [27], Flocchini et al. prove that the pure token model is computationally as powerful as the whiteboard model for the black hole search problem. Furthermore, the complexity is exactly the same. Authors prove that a team of two asynchronous agents, each endowed with a single identical pebble (which can be placed only on nodes, and with no more than one pebble per node) can locate the black hole in an arbitrary network of known topology. This can be done with (n log n) moves, where n is the number of nodes.

IV. External Decontamination- A Classification Review And Evaluation

Due to the manner in which the Network topology is contaminated, decontamination approaches may not be efficiency applied in term of complexity and resource efforts. Therefore we have created a novel classification for decontamination. Our classification space pays particular attention to the impact that the choices of some parameters of the model have on the efficiency of the solutions. In this section we review and compare such recent strategies solution according to three key characteristics: Network topology, effort minimization and complexity results produced.

A. Tree Topology

The tree was the first topology to be investigated in the Decontamination. In [3], Barrière et al. showed that for a given tree *T*, the minimum number of agents needed to decontaminate *T* depends on the location of the homebase. The proposed solution is based on two observations. Consider node *A*, if *A* is not the homebase, the agents will arrive at *A* for the first time from some link *e*. Let $T_1(A), \ldots, T_i(A), \ldots, T_{d(A)-1}$ be the subtrees of *A* from the other incident links, where *d(A)* denotes the degree of *A*, let m_i be the number of agents needed to decontaminate T_i *(A)* once the agents are at *A*, and let $m_i \geq m_{i+1}$, $1 \leq i \leq d(A)-2$. The first observation is that to decontaminate *A* and all its other subtrees without recontamination, the minimum number *m (A)* of agents needed is $m(A) = m_1$ if $m_1 > m_2$ and $m(A) = m_1 + 1$ if $m_1 = m_2$. Consider now homebase *B*, let m_j *(B)* be the minimum number of agents needed to decontaminate the subtree $T_j(B)$, and let $m_j \geq m_{j+1}$, $1 \leq j \leq d(B)$. The second observation is that to decontaminate the entire tree starting from B the minimum number $m(B)$ of agents needed is $m(B) = m_1$ *if* $m_1 > m_2$ *and* $m(B) = m_1 + 1$ *if* $m_1 = m_2$.

Based on these two observations, Barrière et al. [3] first show how the determination of the optimal number of agents can be done through saturation. Simple information about the structure of the tree are collected from the leaves and propagated along the tree, until the optimal number of agents is known for each possible starting point. The most interesting aspect of this strategy is that it immediately yields a protocol for trees that uses the exact minimum number of agents. The technique to determine the minimum number of agents and the corresponding decontamination strategy is done in $O(n)$ time and exchanges $O(n)$ messages. The algorithm is also naturally distributed; the minimum number of agents and the decontamination strategy can be computed in a decentralized manner. The trees that require the largest number of agents are complete binary trees, where the number of agents is $O(\log(n))$. In contrast, in the line two agents are sufficient.

In [29], Flocchini et al. introduce decontamination with *temporal immunity* in a tree. The main difference between the classical decontamination model, and the *temporal immunity* model is that, a cleaner is able to decontaminate any infected node it visits. Once the cleaner departs, the decontaminated node is immune for a certain time $t \geq 0$ (i.e. $t = 0$ corresponds to the model without temporal immunity studied in the previous work) time units to viral attacks from infected neighbors. After the immunity time t is elapsed, recontamination can occur. The minimum team size necessary to disinfect any given tree with immunity time t is derived. Further, Flocchini et al. [29] show how to compute the minimum team size for all nodes of the tree and implicitly the solution strategy starting from each starting node. These computations use a total of $\Theta(n)$ time (serially) or $\Theta(n)$ messages (distributively). Authors then provide a complete structural characterization of the class of trees that can be decontaminated with k agents and immunity time t; Flocchini et al. [29] do so by identifying the forbidden subgraphs and analyzing their properties. Finally, authors consider generic decontamination algorithms, protocols that work unchanged in a large class of trees with little knowledge of their topological structure. Flocchini *et al*. [29] prove that, for each immunity time $t \geq 0$, all trees of a maximum height h can be decontaminated by a team of $k = \left\lfloor \frac{2h}{t+2} \right\rfloor$ agents whose only knowledge of the tree is the bound h.

B. Hypercube Topology

Decontamination in a hypercube has been studied in [25], in which Flocchini et al. prove a lower bound on the number of agents necessary and sufficient to decontaminate a hypercube of size n, $\Theta(\frac{n}{\sqrt{\log n}})$. The employ of this optimal number in the *Local Model* has an interesting consequence, $\Theta(\frac{n}{\sqrt{\log n}})$ is the search number in the classical graph search problem. In the *Local Model* an agent located at a node can see only local information like state of the node, labels of the incident links, and other agent present at the node.

Four different strategies were proposed. In the first strategy (*Local model*), one of the agents acts as a coordinator for the entire cleaning process. The cleaning strategy is carried out on the broadcast tree of the hypercube. The main idea is to place enough agents on the home-base and to have them move, level by level, on the edges of the broadcast tree, led by the coordinator in such a way that no recontamination may occur. This strategy employs an optimal number of agents$\Theta(\frac{n}{\sqrt{\log n}})$, with $O(n \log n)$ moves, and runs in $O(n \log n)$ time steps. The second strategy is devised for a model where the agents are allowed to ``see" the state of their neighbors called *Visibility Model*. Visibility offer to the agent the capability to see whether neighboring node is guarded, clean, or contaminated, in some mobile agent system the visibility power could be easily achieved by probing the state of neighboring node before making a decision.

In this strategy, the computation is local, so there is no need for a coordinator and agents can move autonomously. In fact, the agents are still moving on the broadcast tree, but they do not have to follow the order imposed by the coordinator. In this setting the solution requires $\frac{n}{2}$ agents, but the time complexity is optimal ($\log n$ time steps), and requires the same number of moves $O(n \log n)$. Finally, the last two strategies are devised for models that assume agents have cloning capabilities and can either ``see" the state of their neighbors or move in a synchronous setting. In both cases the bound on the number of moves becomes optimal decreasing to $n - 1$.

C. Mesh Topology

In [28], Flocchini et al. consider the problem of decontaminating a $m \times$ n Mesh $(m \leq n)$. They show some lower bounds on the number of agents, number of moves, and time required to decontaminate an $m \times$ n Mesh $(m \leq n)$. At least m agents, mn moves, and $m + n - 2$ time units are required to solve the decontamination problem. The authors consider two models, one in which an agent has only local knowledge about the node where it resides, and the other in which an agent has visibility to see their neighboring nodes.

In the first model, the only knowledge that an agent has is the information written on the local whiteboard at its current location. The algorithm is described as follows. In the initialization phase, all agents have entered the network in the homebase which is the initial position ($P(0,0)$, upper left most node in the mesh), before the cleaning. Each searching agent is informed by a Synchronizer S to move to its starting point along the first column, but S stays at node $P(0,0)$. By the end of this step there will be one searcher in each node of the first column of the Mesh, while S and one of the searching agents are at the node $P(0,0)$. In the cleaning phase; the Synchronizer S moves *SOUTH* and forces the searching agents to move *EAST* one at a time. When the whole column of agents has moved to the next column, the Synchronizer will also move *EAST* to the next column. Then, it will move *NORTH* and continue to force the searching agents to move *EAST* one at a time. Again, when the whole column of agents has moved to the next column, the Synchronizer will also move *EAST* to the next column. These operations are repeated until the Mesh is cleaned. This algorithm requires $(m+1)$ agents, $\frac{m^2+4mn-5m-2}{2}$ moves, and $(mn-2)$ time units.

In the second model, agents have the power of visibility. Each agent can see the other agents located at its adjacent neighboring nodes and may coordinate its searching operations according to its neighboring agents` moves. In other words, each agent moves independently without the need of a Synchronizer, and agents communicate with each other by using the whiteboard associated with each node in the Mesh. In the initialization phase, all the m agents are located at the node $P(0,0)$ In the cleaning phase, all agents wake up to be the searchers s. Each searcher s will independently perform the following operations: s reads the whiteboard on the current node. If the whiteboard has a ``*CLEAN''* message, $s$ moves *SOUTH* to the next row. $s$ will contiguously move *SOUTH* until it reaches a node on which the Whiteboard is empty. If the whiteboard is empty, $s$ writes a ``*CLEAN''* message on it. s guards the current node until it can see that its neighbouring nodes *NORTH-NORTH*, *NORTH-WEST*, and *NORTH-SOUTH* except *NORTH-EAST* are all clean (or guarded), then s moves *EAST* to the next column. s will repeat this operation until it reaches the last column in the Mesh. The Mesh is cleaned when all the m searching agents reach the last column. This algorithm requires m Agents, $\frac{m^2+2mn-3m}{2}$ moves, and $(m+n-2)$ time unit, time and number of agents complexity are optimal.

In [11], Daadaa *et al.* describe an efficient network decontamination approach. The system consists of a two dimensional lattice that evolves like a cellular automata. A dynamic contamination process causes the spread of a virus (or a fault), and the presence of an agent on a cell guarantees local disinfection (or decontamination). Once disinfected, a cell stays immune to recontamination for a predetermined amount of time. The goal is to design the local rules for the agents and their initial placement so that the agents can decontaminate the entire system without allowing any cell to be re-contaminated. To be efficient, the decontamination should employ as few agents as possible. We design several strategies depending on the type of neighborhood, and on the ability of the agents to clone themselves. The efficiency of the proposed solution depends on the relationship between n and T, where n is the size of the mesh topology and T is time immunity.

D. Tori and Chordal Ring Topologies

In [26], Flocchini et al. studied the decontamination problem in tori and chordal rings. It has been shown that any solution of the decontamination problem in a torus $T(h,k)$ with $h,k \geq 4$ requires at least $2 \times \min(h,k)$ agents, and in the *Local Model* it requires at least $2 \times \min(h,k) + 1$ agents.

To match the lower bound a very simple strategy is employed. The idea is to deploy the agents to cover two consecutive columns and then keep one column of agents to guard from recontamination and have the other column move along the torus. In this setting the solution requires $2h+1$ agents, $hk-2h$ time units and $2hk-4h-1$ moves, where h,k are the dimensions of the torus, $h \leq k$. As for the other topologies, visibility decreases time and slightly increases the number of agents. In the case of torus, it is interesting that in the *Visibility Model* all three complexity measures are optimal. This strategy employs $2h$ agents, $\left\lceil\frac{k-2}{2}\right\rceil$ time units and $hk-2h$ moves. These strategies were generalized to the case of d-dimensional tori.

The *Local* and *Visibility Models* have been also studied in the chordal ring topology. A chordal ring with n nodes is defined as $C(< d_1 = 1, \dots, d_k >)$ and a link structure is defined as $(< d_1 = 1, \dots, d_k >)$ where $d_i < d_{i+1}$ and $d_k \leq \left\lfloor\frac{n}{2}\right\rfloor$. In [26], it is first shown that the smallest number of agents needed for the decontamination does not depend on the size of the chordal ring, but solely on the length of the longest chord. In fact, any solution of the contiguous decontamination problem in a chordal ring $C(< d_1 = 1, \dots, d_k >)$ with $4 \leq d_k \leq \sqrt{n}$ requires at least $2d_k$ agents in the *Local Model* and $d_k + 1$ agents in the *Visibility Model*. In both models, the cleaning is preceded by a deployment stage after which the agents have to occupy $2d_k$ consecutive nodes. After the deployment, the decontamination stage can start. In the *Local Model*, nodes $x0$ to x_{d_k+1} are constantly guarded by one agent each, forming a window of d_k agents. This window of agents will shield the clean nodes from recontamination from one direction of the ring while the agents of the other window are moved by the coordinator (one at a time starting from the one occupying node x_{d_k}) along their longest chord to clean the next window in the ring. Also in the case of the chordal ring, the visibility assumption allows the agents to make their own decision solely on the basis of their local knowledge. An agent move to clean a neighbor only when this is the only contaminated neighbor.

In [36], Luccio et al. studied network decontamination on a k-dimensional torus$(n_1, , n_2, \dots, n_k)$ with $k \geq 1$ and $2 \leq n_1 \leq \cdots \leq n_k$. The decontamination is done by a set of agents moving on a network with local immunity. After an agent leaves from a vertex, this vertex remains uncontaminated as long as m neighbors are uncontaminated. The problem of decontamination is studied for k-dimensional torus with an arbitrary immunity level m, upper and lower bounds are established on the number of agents and of their moves. The proposed approach has required the development of new concepts and algorithms, attaining general results that admit the previous result (where there is no local immunity, $(m = 0)$) is for k-dimensional torus topology as special cases.

V. Conclusion

In this work, we review the exploration in an unsafe environment topic. Two major problems Black-Hole and network decontamination were presented. In the Black-Hole we provide a brief description of synchronous and asynchronous models in order to give the reader a complete knowledge domain. In network decontamination an evaluation methodology was set in terms of a minimum number of system agents needed. We have established both lower and upper bounds in different network topology such as Mesh. Tree, Hypercube, and Chorded ring. The introduction of time immunity will have an important impact on the evaluation results.

ACKNOWLEDGMENT

We would like to thank Paola Flocchini. We would also like to thank the anonymous referees for their useful comments and suggestions.

REFERENCES

[1] S. Albers and M. R. Henzinger. Exploring Unknown Environments. In 29th Annual ACM Symposium on Theory of Computing (STOC), pages 416-425, 1997.

[2] B. Awerbuch, M. Betke, R. L. Rivest, and M. Singh. Piecemeal Graph Exploration by a Mobile Robot. Information and Computation, 152(2): 155-172, 1999.

[3] L. Barrière, P. Flocchini, P. Fraigniaud, and N. Santoro. Capture of an Intruder by Mobile Agents. In ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), pages 200-209, 2002.

[4] L. Barrière, P. Fraigniaud, N. Santoro, and D. M. Thilikos. Searching Is Not Jumping. In 29th International Workshop on Graph Theoretic Concepts in Computer Science (WG), pages 34-45, 2003.

[5] M. A. Bender, A. Fernández, D. Ron, A. Dennunzio Sahai, and S. P. Vadhan. The Power of a Pebble: Exploring and Mapping Directed Graphs. Information and Computation, 176(1):1-21, 2002.

[6] L. Blin, P. Fraigniaud, N. Nisse, and S. Vial. Distributed Chasing of Network Intruders. Theoretical Computer Science, 399(1-2):12-37, 2008.

[7] R. Breish. Intuitive Approach to Speleotopology. Southwestern cavers, 6(5):72-78, 1967.

[8] C. Cooper, R. Klasing, and T. Radzik. Searching for Black-Hole Faults in a Network Using Multiple Agents. In International Conference On Principles Of Distributed Systems (OPODIS), pages 320-332, 2006.

[9] J. Czyzowicz, D. R. Kowalski, E. Markou, and A. Pelc. Complexity of Searching for a Black Hole. Journal Fundamenta Informaticae, 71(2-3):229-242, 2006.

[10] J. Czyzowicz, D. R. Kowalski, E. Markou, and A. Pelc. Searching for a Black Hole in Synchronous Tree Networks. Combinatorics, Probability and Computing, 16(4):595-619, 2007.

[11] Y. Daadaa, P. Flocchini, and N. Zaguia. Decontamination with Temporal Immunity by Mobile Cellular Automata. In International Conference on Scienti_c Computing (CSC), pages 172-178, 2011.

[12] S. Das, P. Flocchini, A. Nayak, and N. Santoro. Distributed Exploration of an Unknown Graph. In 12th International Colloquium on Structural Information Complexity (SIROCCO), pages 99-114, 2005.

[13] X. Deng and C. H. Papadimitriou. Exploring an Unknown Graph (Extended Abstract). In FOCS, pages 355-361, 1990.

[14] K. Diks, P. Fraigniaud, E. Kranakis, and A. Pelc. Tree Exploration with Little Memory. Journal of Algorithms, 51(1):38-63, 2002.

[15] S. Dobrev, P. Flocchini, R. Kralovic, P. Ruzicka, G. Prencipe, and N. Santoro. Black Hole Search in Common Interconnection Networks. Networks, 47(2):61-71, 2006.

[16] S. Dobrev, P. Flocchini, R. Kralovic, and N. Santoro. Exploring an Unknown Graph to Locate a Black Hole Using Tokens. In IFIP TCS, pages 131-150, 2006.

[17] S. Dobrev, P. Flocchini, G. Prencipe, and N. Santoro. Multiple Agents Rendezvous in a Ring in Spite of a Black Hole. In International Conference On Principles Of DIstributed Systems (OPODIS), pages 34-46, 2003.

[18] S. Dobrev, P. Flocchini, G. Prencipe, and N. Santoro. Searching for a Black Hole in Arbitrary Networks: Optimal Mobile Agents Protocols. Distributed Computing,19(1):1-19, 2006.

[19] S. Dobrev, P. Flocchini, G Prencipe, and N. Santoro. Mobile Search for a Black Hole in an Anonymous Ring. Algorithmica, 48(1):67-90, 2007.

[20] S. Dobrev, P. Flocchini, and N. Santoro. Improved Bounds for Optimal Black Hole Search with a Network Map. pages 111-122, 2004.

[21] S. Dobrev, P. Flocchini, and N. Santoro. Cycling Through a Dangerous Network: A SimpleEfficient Strategy for Black Hole Search. In 26th International Conference on Distributed Computing Systems (ICDCS), page 57, 2006.

[22] S. Dobrev, R. Kralovic, N. Santoro, and W. Shi. Black Hole Search in Asynchronous Rings Using Tokens. In 6th Conference on Algorithms and Complexity (CIAC), pages 139-150, 2006.

[23] S. Dobrev, N. Santoro, and W. Shi. Using Scattered Mobile Agents to Locate a Black Hole in an Un-Oriented Ring with Tokens. International Journal of Foundation of Computer Science, 19(6):1355-1372, 2008.

[24] P. Flocchini. Contamination and Decontamination in Majority-Based Systems. Journal of Cellular Automata, 4(3):183-200, 2009.

[25] P. Flocchini, M. J. Huang, and F. L. Luccio. Decontamination of Hypercubes by Mobile Agents. Networks, 52(3):167-178, 2008.

[26] P. Flocchini, M. Jun Huang, and F. L. Luccio. Decontaminating Chordal Rings and Tori using Mobile Agents. International Journal of Foundation of Computer Science, 18(3):547{563, 2007.

[27] P. Flocchini, D. Ilcinkas, and N. Santoro. Ping Pong in Dangerous Graphs: Optimal Black Hole Search with Pure Tokens. In 22nd International Symposium on Distributed Computing (DISC), pages 227-241, 2008.

[28] P. Flocchini, F. L. Luccio, and L. Xiuli Song. Size Optimal Strategies for Capturing an Intruder in Mesh Networks. In Communications in Computing, pages 200-206, 2005.

[29] P. Flocchini, B. Mans, and N. Santoro. Tree Decontamination with Temporary Immunity. In 19th International Symposium on Algorithms and Computation (ISAAC), pages 330-341, 2008.

[30] P. Fraigniaud, L. Gasieniec, D. R.Kowalski, and A. Pelc. Collective Tree Exploration. Networks, 48(3):166-177, 2006.

[31] P. Fraigniaud and N. Nisse. Connected Treewidth and Connected Graph Searching. In 7th Latin American Symposium on Theoretical Informatics (LATIN), pages 479-490, 2006.

[32] P. Glaus. Locating a Black Hole without the Knowledge of Incoming Link. In Algorithmic Aspects of Wireless Sensor Networks (ALGOSENSORS), pages 128-138, 2009.

[33] R. Klasing, E. Markou, T. Radzik, and F. Sarracco. Approximation Bounds for Black Hole Search Problems. In International Conference On Principles Of DIstributed Systems (OPODIS), pages 261-274, 2005.

[34] R. Klasing, E. Markou, T. Radzik, and F. Sarracco. Hardness and Approximation Results for Black Hole Search in Arbitrary Networks. Theoretical Computer Science, 384(2-3):201-221, 2007.

[35] A. Kosowski, A. Navarra, and M. C. Pinotti. Synchronization Helps Robots to Detect Black Holes in Directed Graphs. In International Conference Of Principles Of Distributed Systems (OPODIS), pages 86-98, 2009.

[36] F. Luccio and L. Pagli. A General Approach to Toroidal Mesh Decontamination with Local Immunity. In 2009 IEEE International Symposium on Parallel & Distributed Processing, pages 1-8, 2009.

[37] F. Luccio, L. Pagli, and N. Santoro. Network decontamination in Presence of Local Immunity. International Journal of Foundation of Computer Science, 18(3):457-474, 2007.

[38] P. Panaite and A. Pelc. Exploring Unknown Undirected Graphs. Journal of Algorithms, 33(2):281-295, 1999.

[39] D. Peleg. Size Bounds for Dynamic Monopolies. Discrete Applied Mathematics,86:263-273, 1998.

[40] D. Peleg. Local Majorities, Coalitions and Monopolies in Graphs: a Review. Theoretical Computer Science, 282(2):231-257, 2002.

Semantic Conflicts Reconciliation as a Viable Solution for Semantic Heterogeneity Problems

Walaa S. Ismail
Faculty of Computers and Information, Information Systems Department, Helwan University

Mona M. Nasr
Faculty of Computers and Information, Information Systems Department, Helwan University

Torky I. Sultan
Faculty of Computers and Information, Information Systems Department, Helwan University

Ayman E. Khedr
Faculty of Computers and Information, Information Systems Department, Helwan University

***Abstractt*—Achieving semantic interoperability is a current challenge in the field of data integration in order to bridge semantic conflicts occurring when the participating sources and receivers use different or implicit data assumptions. Providing a framework that automatically detects and resolves semantic conflicts is considered as a daunting task for many reasons, it should preserve the local autonomy of the integrated sources, as well as provides a standard query language for accessing the integrated data on a global basis. Many existing traditional and ontology-based approaches have tried to achieve semantic interoperability, but they have certain drawbacks that make them inappropriate for integrating data from a large number of participating sources.**

We propose semantic conflicts reconciliation (SCR) framework, it is ontology-based system in which all data semantics explicitly described in the knowledge representation phase and automatically taken into account through the interpretation mediation service phase, so conflicts detected and resolved automatically at the query time.

***Keywords—Data Integration; Heterogeneous Sources; Interoperability; Semantic Conflicts; Context; Reconciliation Ontology*.**

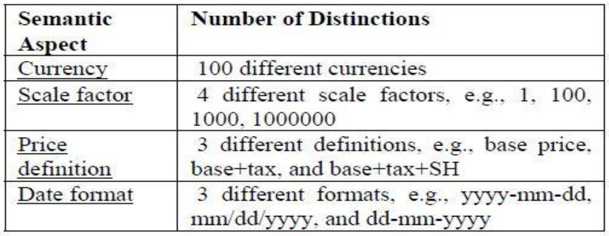
I. INTRODUCTION

Despite the fact that a typical large organization spends nearly 30% of its IT budget on integration and interoperation related efforts, many inter- and intra- organizational systems still have poor interoperability [10]. Technologies already exist to overcome the heterogeneity in hardware, software, and syntax that is used in different systems (e.g., the ODBC standard, XML based standards, web services and SOA-Service Oriented Architectures) .While these capabilities are essential to information integration, they do not address the issue of heterogeneous data semantics that exist both within and across enterprises [11].

Heterogeneity problem occurs when data sources and receivers use different contexts (assumptions); a user submit query and interprets the results in a certain context, which completely different from contexts received from sources. Implicit assumptions made in each source need to be explicitly described and used to reconcile conflicts when data from these systems are combined [3]. Ontology plays an important role on making domain assumptions unambiguous or uniquely identifies the meaning of concepts in a specific domain of interest.

Let us assume that the comparison service covers 100 countries, each having its unique currency and each consisting of 100 vendors. Thus, there are a total of 10,000 sources in this example. For simplicity, let's assume the consumer chooses his context to be the same as one of the sources. Although all vendors in the same country may use the same currency for price, they may use different price definitions and scale factors [9]. Table 1 summarizes the potential context differences in terms of just these four semantic aspects : currency, scale factor, price definition, and date format (for the purpose of finding exchange rate at a given day).

TABLE I. Semantic Differences in Data Sources [9]



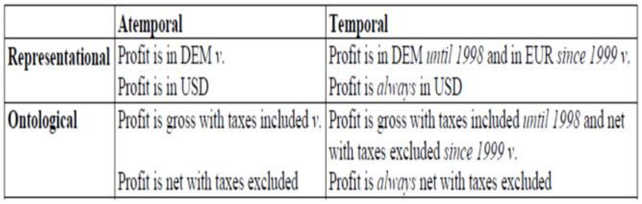
Semantic Aspect	Number of Distinctions
Currency	100 different currencies
Scale factor	4 different scale factors, e.g., 1, 100, 1000, 1000000
Price definition	3 different definitions, e.g., base price, base+tax, and base+tax+SH
Date format	3 different formats, e.g., yyyy-mm-dd, mm/dd/yyyy, and dd-mm-yyyy

Thus, there could be 3600 (i.e., 100*4*3*3) different contexts amongst these sources; e.g., one source has US dollars for currency, scale factor being 1, price as tax and shipping and handling included, with mm/dd/yyyy date format; another source has Turkish liras for currency, scale factor being 1000000, price as only tax included, with dd-mm-yyyy date format, etc. The online comparison service needs to implement the conversions so that the comparison can be performed for sources in any context.

Implementing tens of thousands of data conversions is not an easy task; but maintaining them to cope with changes in data sources and receiver requirements over time is even more challenging [2]

According to Firat [1] there are three dimensions of semantic heterogeneity: contextual, ontological and temporal. Contextual heterogeneity occurs when different systems (sender/receiver) make different assumptions about the representation of the same concept, such as the profit of a company can be represented in DEM (i.e., Deutschmarks) in one system or in USD (i.e., U.S. dollars) in another, where the currency used is the assumption. So there will be two or more not identical representations of the same thing. Ontological heterogeneity occurs when different meanings denoted by the same term (e.g., whether the profit is gross profit including taxes or net profit excluding taxes) because there is a definitional conflicts concerning the inclusion or exclusion of

TABLE II. Temporal vs. Atemporal heterogeneity [4].tax in the profit.



	Atemporal	Temporal
Representational	Profit is in DEM *v.* Profit is in USD	Profit is in DEM *until 1998* and in EUR *since 1999 v.* Profit is *always* in USD
Ontological	Profit is gross with taxes included *v.* Profit is net with taxes excluded	Profit is gross with taxes included *until 1998* and net with taxes excluded *since 1999 v.* Profit is *always* net with taxes excluded

Both the representational and the ontological assumptions can be static and do not change over time within an interested time period, in which case time is not of concern. The resulting heterogeneity is atemporal. Conversely, the assumptions can change over time, and the resulting heterogeneity is temporal [4].

There should be systematic approaches in order to reconcile semantic heterogeneity among heterogeneous sources and receivers.

II. Existing Approaches For Achieving Semantic Interoperability

We can resolve semantic conflicts by hand-coded programs but on small scale only; alternative solutions are needed as the number of systems and the complexity of each system increase.

A. *Traditional Approaches*

Brute-force Data Conversions (*BF*)

In the Brute-force Data Conversions (BF) approach all necessary conversions implemented with hand-coded programs. For example, if we have N data sources and receivers, N (N-1) such conversions need to be implemented to convert the sources context to the receiver context. These conversions become costly to implement and very difficult to maintain When N is large. This is a labor-intensive process; nearly 70% of integration costs come from the implementation of these data conversion programs. A possible variation of the (BF) approach is to group sources that share the same set of semantic assumptions into one context. The approach allows multiple sources in the same context to share the same conversion programs, so the numbers of conversion programs will be reduced. We refer to the original approach and this variation as BFS and BFC, respectively [2]. These approaches are illustrated schematically in Fig 1.

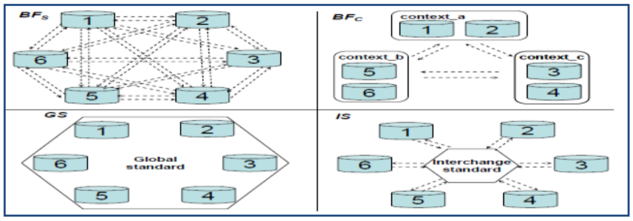


Fig.1. Traditional approaches to Semantic Interoperability [9].

Global Data Standardization (GS)

If we could develop and maintain a single data standard that defines a set of concepts and specifies the corresponding representation, all semantic differences would disappear and there would be no need for data conversion. Unfortunately, such standardization is usually infeasible in practice for several reasons. There are legitimate needs for having different definitions for concepts, storing and reporting data in different formats. Most integration and information exchange efforts involve many existing systems, agreeing to a standard often means someone has to change his/her current implementation, which creates obstacles and makes the standard development and enforcement extremely difficult [7].

Interchange Data Standardization (IS)

Data exchange systems can sometimes agree on the data to be exchanged, i.e., standardizing a set of concepts as well as their interchange formats. The underlying systems do not need to store the data according to the standard; it suffices as long as each data sender generates the data according to the standard. That is, this approach requires that each system have conversions between its local data and an interchange standard used for exchanging data with other systems. Thus, each system still maintains its own autonomy. This is different from the global data standardization, where all systems must store data according to a global standard. With N systems exchanging information, the Interchange Standardization approach requires 2N conversions. The IS approach is a significant improvement over the brute-force approach that might need to implement conversions between every pair of systems [9]. Although this approach has certain advantages, it also has several serious limitations [2]. From which, all parties should reach an agreement on the data definition and data format. Reaching such an agreement can be a costly and time-consuming process besides; any change to the interchange standard affects all systems and the existing conversion programs. Lastly, the approach can involve many unnecessary data conversions

B. *Ontology-Based Data Integration Approaches*

Most of the shortcomings in the previous traditional approaches can be overcome by using ontology-based systems .We explain the most popular ontology-based systems for data integration, which are SCROL and COIN with respect to the role and use of ontologies.

SCROL is a global schema approach that uses an ontology to explicitly categorize and represent predetermined types of semantic heterogeneity [6]. It is based on the use of a common ontology, which specifies a vocabulary to describe and interpret shared information among its users. It is similar to the federated schema approach. However, an ontology-based domain model captures much richer semantics and covers a much broader range of knowledge within a target domain. But it uses a fully specified ontology to explicitly categorize and represent predetermined types of semantic heterogeneity. SCROL assumes that the underlying information sources are structured data that may reside in the structurally organized text files or database systems. However, the unprecedented growth of Internet technologies has made vast amounts of resources instantly accessible to various users via the World Wide Web (WWW) [6].

COIN Project was initiated in 1991 with the goal of achieving semantics interoperability among heterogeneous information sources. The main elements of this architecture are wrappers, context axioms, elevation axioms, a domain model, context mediators, an optimizer and an executioner. A domain model in COIN is a collection of primitive types and semantic types (similar to type in the object-oriented paradigm), which defines the application domain corresponding to the data sources that are to be integrated COIN introduces a new definition for describing things in the world. It states that the truth of a statement can only be understood with reference to a given context. The context information can be obtained by examining the data environment of each data source [11].

The problem of semantic interoperability is not new, and people have tried to achieve semantic interoperability in the past using various approaches. Traditional approaches have sometimes been reasonably successful in limited applications, but have proven either very costly to use, hard to scale to larger applications, or both. Traditional approaches have certain drawbacks that make them inappropriate for integrating information from a large number of data sources. Existing ontology-based approaches for semantic interoperability also have not been sufficiently effective because there is no systematic methodology to follow, no concert methodology for building ontologies and all existing ontology-based not able to reconcile all types of semantic conflicts.

III. SCR Architecture

The Semantic Conflicts Reconciliation (SCR) framework is considered as ontology based system aims to solve semantic data level conflicts among different sources and receivers in a systematic methodology. SCR is based on domain specific ontology to create user queries. The user can browse the merged ontology and selects specific terms and conditions to create global query. There is no need for the user to be aware of terms in databases in order to query them. The selected terms are mapped to the corresponding terms in each data source to decompose the global query to a set of sub naïve queries. The decomposed sub-queries are converted to well-formed sub-queries before sending it to the suitable database. Finally the SCR combine and resend the well-formed query results after reconciling the detected conflicts to the users according to the required contexts.

SCR consists of two phases, the knowledge representation phase and the interpretation mediation service phase [8].

A. Knowledge Representation

The knowledge representation phase consists of the following components:

- Ontology Extraction: Extract local ontology from each database.
- Global Ontology: Merge all local ontologies to construct a global one that contains all major concepts and the relationships between them.
- Contexts: Explicitly describing the sources and receivers assumptions about data.
- Mapping: Linking between the constructed merged ontology and the corresponding terms in each data source in order to produce the semantic catalog.

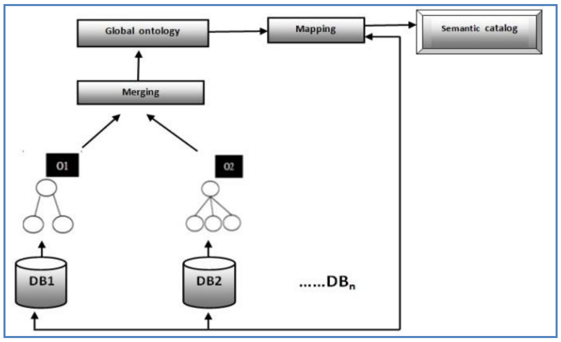


Fig.2. Knowledge representation phase[8].

Database to Ontology Extraction:

In the Ontology extraction step, we have multiple databases to extract a local ontology from each one. A local ontology contains all database information like tables, columns, relations, constraints. Moreover, it contains intentional definitions to represent higher level of abstraction than traditional data models.

The local ontology represents a relational database tables as concept and columns as slots of the concept. The local ontologies are represented in a formal standard language called OWL (Ontology Web Language).

Creating local ontology for each database saves them independent. Any changes in the schema or relations can be added easily to its local ontology. The local ontology includes only the metadata and additional semantics; however, the database instances or members still in the data sources separated from its ontology.

Global Ontology Construction:

Our framework is based on the hybrid ontology approach in which we create local ontology for each data source and a global ontology which is considered a reference for all local ontologies involved in the integration process.

The Merging process aims to create one global (merged) ontology that contains multiple local ontologies contents. It contains all the knowledge of the initial ontologies [5].in order to create a merged ontology, the corresponding objects will be matched from two or more local ontologies. Subsequently, suitable matching algorithm should choose. Matching is the core of the merging process to make one vantage point of view from multiple ontologies, where some concepts and slots will be represented as a new concept and new slots, or some slots may be merged and follow another concept. We can say that, there is a new structure that will be created in the merged ontology. This structure does not affect the information sources, because each local ontology is independent. Creating a standard formal model (merged ontology) makes query multiple databases satisfy the user requirements at the semantic level.

The SCR framework uses PROMPT tool to matching and merging local ontologies. PROMPT is a semi-automatic tool. It is protégé plug in. It guides the expert by providing suggestions. PROMPT provides suggestions about merging and copying classes. Figure 3 explains the PROMPT algorithm [4]. PROMPT takes two ontologies as input and guide the user to create a merged ontology as output. IT generates a list of suggestions based on the choose matching algorithm. Our Framework uses PROMPT lexical matching algorithm.

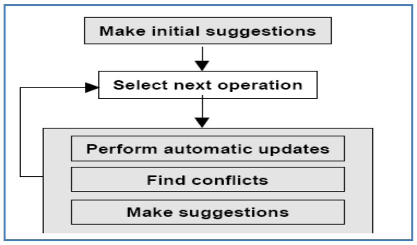


Fig.3. The Flow of PROMPT Algorithm [4].

According to Noy and Musen [4] about PROMPT evaluation, human experts follow 90% of PROMPT suggestions. During the merging process, PROMPT suggested 74% of the total knowledge_ base operations that the user invoked. PROMPT s able to perform a large number of merging operations on its own (or with simple "approval" of a human expert). Thus, it can save the expert’s time and efforts.

Explicitly define contexts

The proposed framework assigns contexts descriptions about data items in each source using the following two steps.

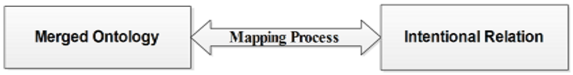
Adding annotation: Adding annotation properties (modifiers) to the global ontology slots to denote their contexts. We consider annotation properties as special properties that affect the interpretation of data values.

Value assignment: Assign values that explicitly describe the semantics of data in different aspects for each annotation properties created in the previous step.

We can associate more than one annotation property to the same. We can easily add, remove and change the assigned values in the ontology whenever the context changed in the sources over time.

Mapping global ontology to multiple databases:

We developed a semi automatic mapping tool used to map the merged ontology to multiple databases. The main purpose of the proposed mapping tool is to find and match between semantically similar terms in the global query with the corresponding terms in the data sources of the integrated system. The output of the mapping process is the semantic catalog.



The mapping process in our proposed mapping tool follows the following steps:

- The mapping process started by creating a database with two tables, to save the mapping data in the first table, and saving the metadata of the database system in the second table. This process is done once when the mapping process is started.
- The expert selects the first database in the heterogeneous integrated sources to link its schema (intentional relation) with the terms in the global ontology.
- When the user selects database from a list of all databases that existed, then all tables in the selected database will be listed. Then, press to select columns, all columns in the selected table will be listed and saved in the table created in the first step along with the correspondence terms in the global ontology. All the primary keys, foreign keys, and referenced tables for each table in the selected database are automatically retrieved and saved in the second created table as metadata, to use it in query processing.

B. *Interpretation Mediation Service*

Knowledge representation phase is not enough to solve semantic conflicts among data sources and receivers. The second phase in our proposed system is the interpretation mediation service in which the user interacts with the system through graphical user interface (GUI). With the Interpretation Mediation Service support the user no longer concerned about context differences and how contexts evolve.

The SCR framework architecture, consisting of the following four main components.

- System interface
- Query preprocessor
- SQL generator
- Converter (query engine)

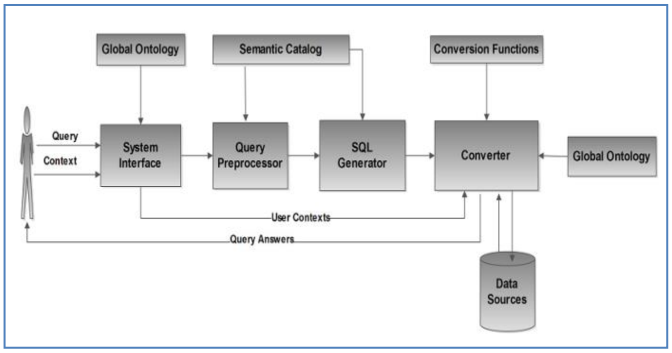


Fig.4. Architecture of the SCR Framework.

System Interface

We cannot suppose that users have intimate knowledge about data sources being queried especially when the number of these sources are big. Users should remain isolated from semantic conflicts problems and no need to learn one of the ontology query languages to formulate his query. User interacts with the system through graphical user interface (GUI). The GUI displays the global ontology terms to facilitate finding the global query terms easily and quickly. User browses the global ontology to select specific terms for his query.

Query preprocessor

Query preprocessor receives the global query terms and semantic catalog as input and produce blocks of user's data based on the selected items from the system interface. Each block represents a query but without any language format. Once the user selects terms and conditions from the system the query preprocessor does the following actions.

- Query preprocessor utilizes semantic catalog (mapping file) to retrieve the database name, table name and columns names that mapped to the selected terms and conditions in the user query.
- The query preprocessor reorganizes the retrieved data from the previous step into blocks according to the database name. Each block represents a query but it does not present in any language format.

SQL Generator

SQL generator turns the query blocks received from the query preprocessor into SQL queries and directs them to the converter. It uses the semantic catalog (metadata) to translate the previous blocks into SQL correct syntax. In order to transform the blocks to correct syntax, the generator adds select, from and where clauses. In addition, if the query needs to retrieve instances from more than one table the primary keys, foreign keys and referenced tables of the integrated databases may be added from the semantic catalog metadata file as well.

Converter (query engine)

We consider converter as a query engine that takes SQL queries from the SQL generator and the user context as input. Converter connects to the merged (global) ontology to retrieve and compare annotations in order to transform the user naïve query (that ignores differences in assumptions between sources) into a well-formed query that respects differences among sources and receivers contexts. The query engine rewrites the user query into a mediated query (multiple sub-queries) with a set of instructions and functions in order to reconcile the semantic conflicts.

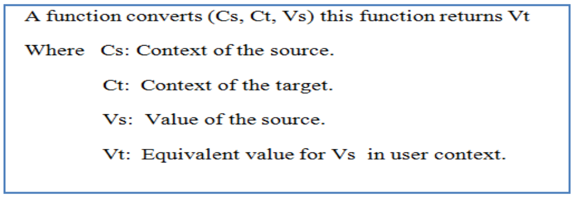
The Converter detects and reconciles semantic conflicts between sources and receivers according to the following two stages:

First stage: before sending the SQL queries to the suitable data source, the query engine connects to the merged ontology and compares the annotation values between the global query contexts and the data sources contexts that are involved in the global query. The query engine detects and reconciles the conflicts at the query time then directs each SQL query to the suitable database.

Second stage: the query engine compares the annotation values of each item in the sources involved in the global query with the user required context .

Whether in the first or the second stage, the converter connects to a set of conversion functions defined in order to convert between contexts and satisfy the user expectations about results.

The general form of the conversion function can be as follows:



A function converts (Cs, Ct, Vs) this function returns Vt

Where Cs: Context of the source.

Ct: Context of the target.

Vs: Value of the source.

Vt: Equivalent value for Vs in user context.

Fig.5. General form of the conversion functions[8].

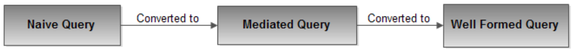
Conversion functions represent the conversions among all annotation property values or contexts described in the merged ontology. In the SCR system, there is no relation between the number of sources or receivers and the number of conversion functions. Conversion functions in SCR system are parameterized Component conversion that is defined for each modifier in the ontology. These functions can convert between all values of a modifier automatically (e.g., a conversion that reconciles between currencies of price modifier).

IV. SCR Software System Description

We developed the SCR software system for demonstrating the feasibility and the features of our approach. It helps the user to query integrated sources with minimal efforts using Query-by-Example language (QBE). As a result, the user needs only to know little information about the global ontology terms.

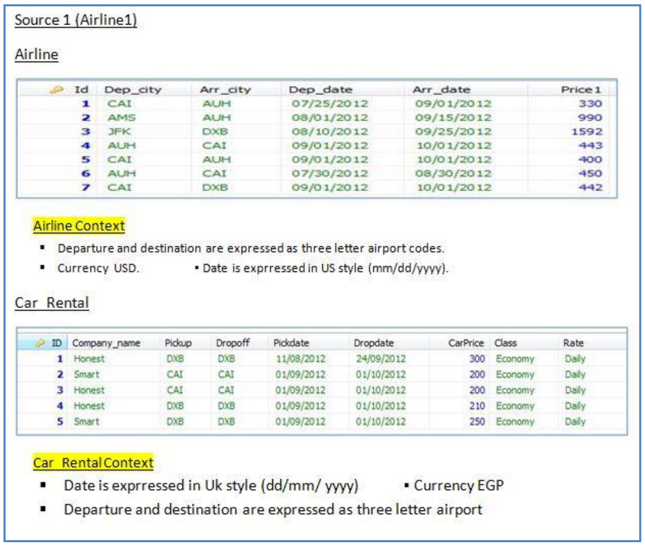
A. *Naive to well-formed query conversion*

Consider the following scenario in order to describe how the SCR system transforms the user naïve query into a well-formed query.



Illustrative example (Airline Reservation Scenario)

Consider the example where a comprehensive travel system involving multiple airlines, and car rental services. Assume we have two data sources and each source has its own contexts. The User can find the most suitable and best cheapest airline booking along with the cheapest car rental prices through the different car rental providers available. The airline reservation scenario displays prices and airline information from different airlines data sources given departure and destination locations and dates. We show snapshots from this scenario in Figure 6 and Figure 7. As shown there are many conflicts between the sources contexts in addition to the user different contexts or assumptions about data.



Source 1 (Airline1)

Airline

Id	Dep_city	Arr_city	Dep_date	Arr_date	Price 1
1	CAI	AUH	07/25/2012	09/01/2012	330
2	AMS	AUH	08/01/2012	09/15/2012	990
3	JFK	DXB	08/10/2012	09/25/2012	1592
4	AUH	CAI	09/01/2012	10/01/2012	443
5	CAI	AUH	09/01/2012	10/01/2012	400
6	AUH	CAI	07/30/2012	08/30/2012	450
7	CAI	DXB	09/01/2012	10/01/2012	442

Airline Context

- Departure and destination are expressed as three letter airport codes.
- Currency USD. ▪ Date is exprressed in US style (mm/dd/yyyy).

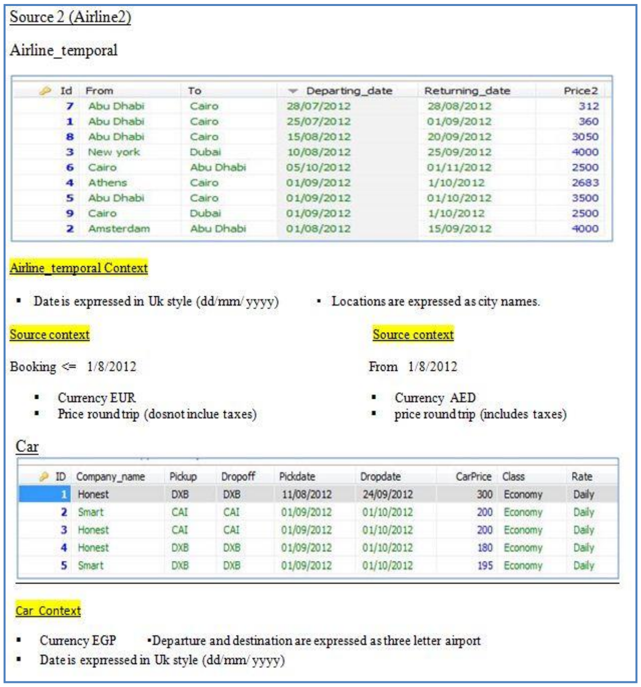
Car Rental

ID	Company_name	Pickup	Dropoff	Pickdate	Dropdate	CarPrice	Class	Rate
1	Honest	DXB	DXB	11/08/2012	24/09/2012	300	Economy	Daily
2	Smart	CAI	CAI	01/09/2012	01/10/2012	200	Economy	Daily
3	Honest	CAI	CAI	01/09/2012	01/10/2012	200	Economy	Daily
4	Honest	DXB	DXB	01/09/2012	01/10/2012	210	Economy	Daily
5	Smart	DXB	DXB	01/09/2012	01/10/2012	250	Economy	Daily

Car Rental Context

- Date is exprressed in Uk style (dd/mm/ yyyy) ▪ Currency EGP
- Departure and destination are expressed as three letter airport

Fig.6. Data assumptions in source1 (Airline1)



Source 2 (Airline2)

Airline_temporal

Id	From	To	Departing_date	Returning_date	Price2
7	Abu Dhabi	Cairo	28/07/2012	28/08/2012	312
1	Abu Dhabi	Cairo	25/07/2012	01/09/2012	360
8	Abu Dhabi	Cairo	15/08/2012	20/09/2012	3050
3	New york	Dubai	10/08/2012	25/09/2012	4000
6	Cairo	Abu Dhabi	05/10/2012	01/11/2012	2500
4	Athens	Cairo	01/09/2012	1/10/2012	2683
5	Abu Dhabi	Cairo	01/09/2012	01/10/2012	3500
9	Cairo	Dubai	01/09/2012	1/10/2012	2500
2	Amsterdam	Abu Dhabi	01/08/2012	15/09/2012	4000

Airline_temporal Context

- Date is exprressed in Uk style (dd/mm/ yyyy) ▪ Locations are expressed as city names.

Source context

Booking <= 1/8/2012

- Currency EUR
- Price round trip (dosnot inclue taxes)

Source context

From 1/8/2012

- Currency AED
- price round trip (includes taxes)

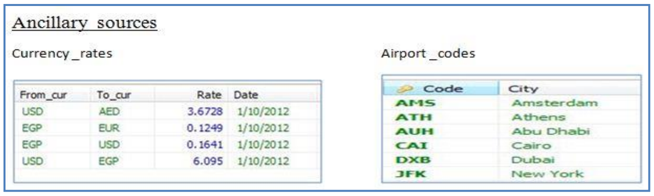
Car

ID	Company_name	Pickup	Dropoff	Pickdate	Dropdate	CarPrice	Class	Rate
1	Honest	DXB	DXB	11/08/2012	24/09/2012	300	Economy	Daily
2	Smart	CAI	CAI	01/09/2012	01/10/2012	200	Economy	Daily
3	Honest	CAI	CAI	01/09/2012	01/10/2012	200	Economy	Daily
4	Honest	DXB	DXB	01/09/2012	01/10/2012	180	Economy	Daily
5	Smart	DXB	DXB	01/09/2012	01/10/2012	195	Economy	Daily

Car Context

- Currency EGP ▪Departure and destination are expressed as three letter airport
- Date is exprressed in Uk style (dd/mm/ yyyy)

Fig.7. Data Assumptions in Source2 (Airline2)



Ancillary sources

Currency_rates

From_cur	To_cur	Rate	Date
USD	AED	3.6728	1/10/2012
EGP	EUR	0.1249	1/10/2012
EGP	USD	0.1641	1/10/2012
USD	EGP	6.095	1/10/2012

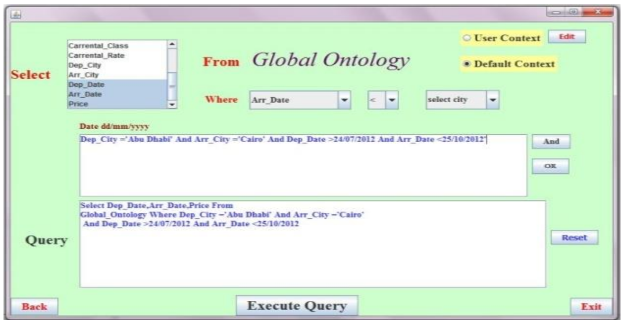
Airport_codes

Code	City
AMS	Amsterdam
ATH	Athens
AUH	Abu Dhabi
CAI	Cairo
DXB	Dubai
JFK	New York

Fig.8. Ancillary Tables

According to Firat [1] we mentioned that there are three dimensions of semantic heterogeneity: contextual, ontological and temporal. Now we describe how our proposed framework (SCR) can reconcile these conflicts.

Suppose the user submit the following global query1 to the SCR system in order to know the available airlines along with their prices.

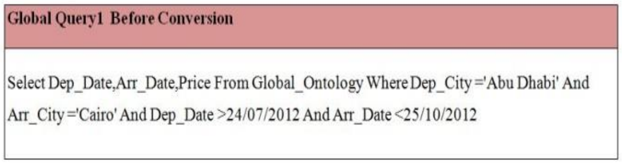


Dep_City ='Abu Dhabi' And Arr_City ='Cairo' And Dep_Date >24/07/2012 And Arr_Date <25/10/2012'

Select Dep_Date,Arr_Date,Price From
Global_Ontology Where Dep_City ='Abu Dhabi' And Arr_City ='Cairo'
And Dep_Date >24/07/2012 And Arr_Date <25/10/2012

Fig.9. Global query1

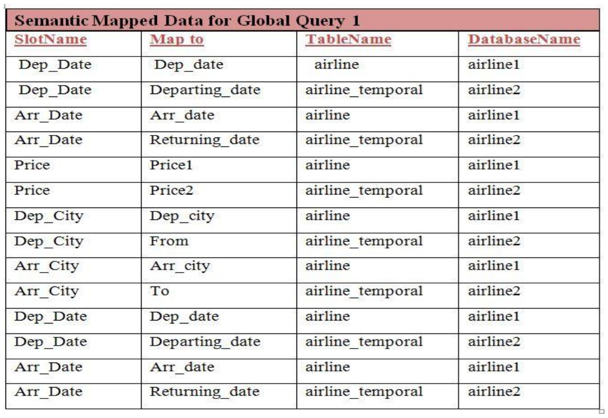
1. The GUI displays the global ontology terms to facilitate finding the global query terms easily and quickly. User browses the global ontology to selects specific terms and conditions for query1 as shown in Figure 9.



Global Query1 Before Conversion

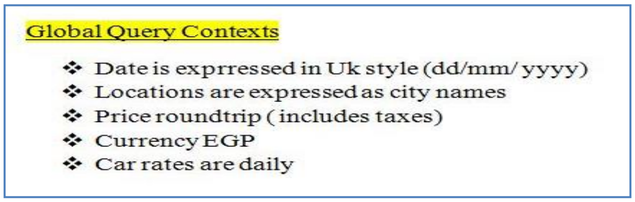
Select Dep_Date,Arr_Date,Price From Global_Ontology Where Dep_City ='Abu Dhabi' And Arr_City ='Cairo' And Dep_Date >24/07/2012 And Arr_Date <25/10/2012

2. Query preprocessor utilizes semantic catalog (mapping file) to retrieve the semantic mapped data for global query1 that consists of database name, table name and columns names that mapped to each selected term and condition in the user global query1 and reorganizes the retrieved data into blocks according to the database name.



Semantic Mapped Data for Global Query 1			
SlotName	Map to	TableName	DatabaseName
Dep_Date	Dep_date	airline	airline1
Dep_Date	Departing_date	airline_temporal	airline2
Arr_Date	Arr_date	airline	airline1
Arr_Date	Returning_date	airline_temporal	airline2
Price	Price1	airline	airline1
Price	Price2	airline_temporal	airline2
Dep_City	Dep_city	airline	airline1
Dep_City	From	airline_temporal	airline2
Arr_City	Arr_city	airline	airline1
Arr_City	To	airline_temporal	airline2
Dep_Date	Dep_date	airline	airline1
Dep_Date	Departing_date	airline_temporal	airline2
Arr_Date	Arr_date	airline	airline1
Arr_Date	Returning_date	airline_temporal	airline2

3. SQL Generator uses the semantic catalog (metadata) to translate the previous created blocks received from the query preprocessor into SQL correct syntax.
4. The created queries are named a naïve queries, because the direct execution of them would not respect the semantic conflicts and would most likely return empty or semantically incorrect answers. If the created queries from the global query1 submitted to Airline1 and Airline2 databases without any mediation (conversion) would return empty results from source1 and semantically incorrect result set from source2.
5. Before sending the queries to the suitable data source the query engine connects to the merged ontology and compares the annotation values between the global query context (Figure 10) and the data sources contexts (Figure 6 and Figure 7) that involved in global query1.



Global Query Contexts

- Date is exprressed in Uk style (dd/mm/ yyyy)
- Locations are expressed as city names
- Price roundtrip (includes taxes)
- Currency EGP
- Car rates are daily

Fig.10. The Global Query Context

6. The SCR query engine detects the following semantic conflicts

Contextual Heterogeneity: As we mentioned before, Contextual heterogeneity occurs when different systems (sender/receiver) make different assumptions about the representation of the same concept) so there will be two or more not identical representations of the same thing. Such as in query 1 there are different representation of date format (UK vs. US Date Format) and locations (city name vs. airport code). The query engine detects two Contextual conflicts between the created SQL query contexts (Figure 10) and source 1 contexts:

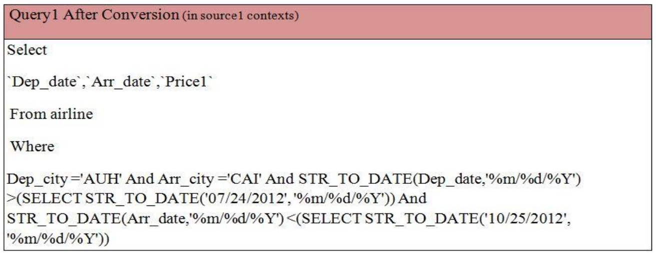
Date is expressed in US style (mm/dd/yyyy) in source 1 (Airline) and in Uk style (dd/mm/ yyyy) in the query context. This type of conflict can be solved in the mediation step using direct conversion from UK to US .

Cities represented as three letters airport code in source 1 (Airline) and as city full names in the query context. This type of conflict requires auxiliary tables in order to reconcile them as shown in Figure 8.

7. The converter connects to the conversion functions in order to reconcile the previous contextual conflicts at the query time by mediated queries then directs each query to the suitable database.

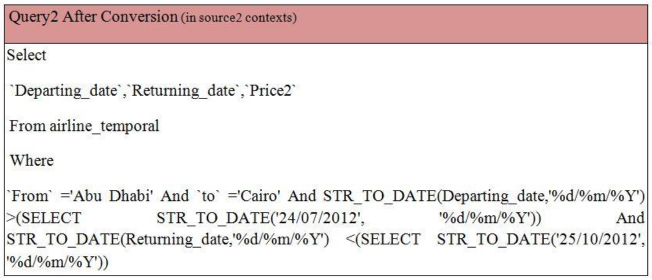
City name conflicts were dynamically reconciled, with the help of Airport_codes table that is used as ancillary table to convert from full city names to airport codes.

Date conflicts were statically reconciled by converting date values from Uk style (dd/mm/ yyyy) to US style (mm/dd/yyyy) using the direct conversion operation.



Query1 After Conversion (in source1 contexts)

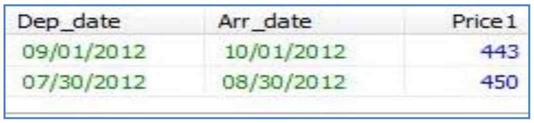
```
Select
`Dep_date`,`Arr_date`,`Price1`
From airline
Where
Dep_city ='AUH' And Arr_city ='CAI' And STR_TO_DATE(Dep_date,'%m/%d/%Y')
>(SELECT STR_TO_DATE('07/24/2012', '%m/%d/%Y')) And
STR_TO_DATE(Arr_date,'%m/%d/%Y') <(SELECT STR_TO_DATE('10/25/2012',
'%m/%d/%Y'))
```



Query2 After Conversion (in source2 contexts)

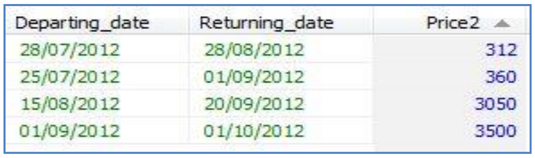
```
Select
`Departing_date`,`Returning_date`,`Price2`
From airline_temporal
Where
`From` ='Abu Dhabi' And `to` ='Cairo' And STR_TO_DATE(Departing_date,'%d/%m/%Y')
>(SELECT STR_TO_DATE('24/07/2012', '%d/%m/%Y')) And
STR_TO_DATE(Returning_date,'%d/%m/%Y') <(SELECT STR_TO_DATE('25/10/2012',
'%d/%m/%Y'))
```

8. Now If queries after the previous conversion submitted to the suitable database would return semantically incorrect results from source 1 (Figure 6) and from source 2 (Figure 7) because the retrieved results not respect the user expectations (user required context).



Dep_date	Arr_date	Price 1
09/01/2012	10/01/2012	443
07/30/2012	08/30/2012	450

Fig.11. Source1 Semantically Incorrect Results of Query1

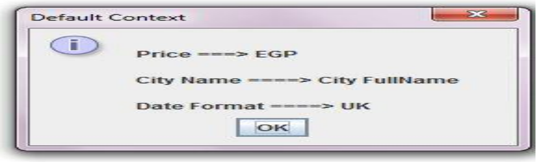


Departing_date	Returning_date	Price2 ▲
28/07/2012	28/08/2012	312
25/07/2012	01/09/2012	360
15/08/2012	20/09/2012	3050
01/09/2012	01/10/2012	3500

Fig.12. Source2 Semantically Incorrect Results of Query1

9. After directing each query to the suitable data source, then query engine compares the annotation values of each item in the sources with the user required contexts.

In global query 1 our user selects to display the query results in the default context.



Default Context

Price ===> EGP

City Name ====> City FullName

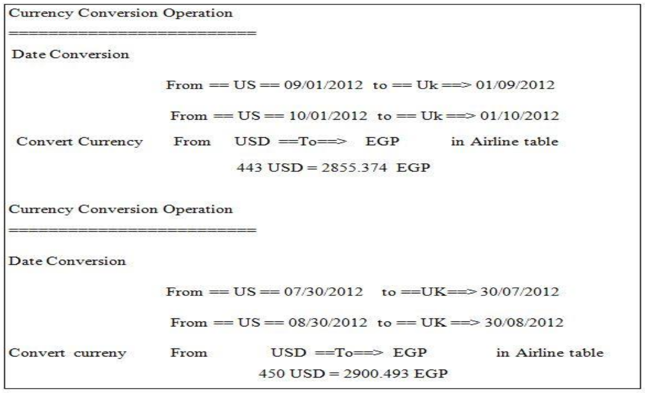
Date Format ====> UK

Query1 Semantic Conflicts (Airline1):

The query engine detects two contextual conflicts between the user contexts (default) and source1 contexts

Contextual conflict in the price concept, currency modifier value in source1 is USD context but in EGP in the user context. Such conflict detected and reconciled through the mediated query. Date is expressed in US style (mm/dd/yyyy) in source1 (Airline) but our user expects it as UK style (dd/mm/ yyyy).

The previous contextual conflicts reconciled on the query time by the converter through the mediated queries. Figure 13 shows a Trace of reconciling source 1 detected contextual semantic conflict as follow.



```
Currency Conversion Operation
=========================
 Date Conversion

                              From == US == 09/01/2012  to == Uk ==> 01/09/2012

                               From == US == 10/01/2012  to == Uk ==> 01/10/2012
  Convert Currency      From     USD  ==To==>    EGP          in Airline table
                                        443 USD = 2855.374  EGP

Currency Conversion Operation
=========================

Date Conversion

                              From == US == 07/30/2012    to ==UK==> 30/07/2012

                               From == US == 08/30/2012  to == UK ==> 30/08/2012
Convert  curreny          From            USD  ==To==>  EGP            in Airline table
                                              450 USD = 2900.493 EGP
```

Fig.13. Trace of reconciling source1 contextual semantic conflict

Query 1 Semantic Conflicts (Airline 2):

The query engine detects semantic mismatches among contexts caused by the implicit assumptions. The implicit assumptions not only differ between the two sources, but also changed in the same source from one context to another over time as shown in the airline_temporal table in Figure 7.
The user expects the price always be in EGP and with taxes included. While in source 2 (airline_temporal table) the price is in EUR and excluding taxes if departing date <= 01/08/2012 and changed the currency from EUR to AED and price includes taxes if departing date > 01/08/2012 , Such conflicts lead us to the third dimension of semantic heterogeneity which is the temporal dimension).

A. Reconciling Temporal Semantic Heterogeneity:

Temporal semantic heterogeneities are related to both contextual and ontological heterogeneities and occur in situation where the semantics between data sources, even in the same data source, change over time [11]. Both the representational and the ontological assumptions can be static and do not change over time within an interested time period, in which case time is not of concern (same previous examples) or, the assumptions can change over time, and the resulting heterogeneity is temporal. When the implicit assumptions change over time, data corresponding to different time periods are subject to different interpretations. Based on the previous definition we have two categories of temporal heterogeneity, temporal representational (contextual) heterogeneity and temporal ontological heterogeneity.

It's challenging to deal with data if it's meaning changes overtime. These challenges will be more difficult if we want to handle the changes through the integration of multiple heterogeneous sources. We describe the implicit assumptions about data elements in temporal concepts by similar way of describing non temporal concepts by adding annotations (modifiers) in the merged ontology for each term.
We identify the following approach to explicitly describe the temporal assumptions in the knowledge representation phase and reconciling the temporal semantic conflicts through the interpretation mediation service phase.
Proposed Approach for Representing Temporal Contexts
Temporal assumptions approach: Describing each data element (attribute) in a data source that has assumptions change over time with different ontological concepts at different times. So we convert temporal assumptions to set of static assumptions within an interested time period.

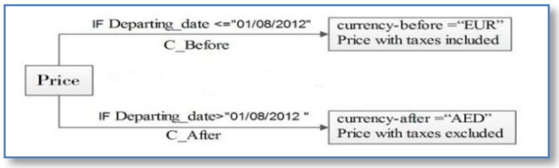

Price
IF Departing_date <="01/08/2012"
C_Before
currency-before ="EUR"
Price with taxes included
IF Departing_date>"01/08/2012 "
C_After
currency-after ="AED"
Price with taxes excluded

Fig.14. Temporal Assumptions Approach

In query 1 we explain the representation of temporal contexts in the ontology according to our proposed approach as shown in Figure 14.

Reconciling Temporal Representational Conflicts of Query1

In temporal representational assumptions the same attribute may be represented differently at different times of a data source according to our previous approach.
In order to resolve the temporal contextual conflict between our user context of query 1 and the source 2 context of airline_temporal table as shown in Figure 7. We explicitly describe the price data element in the ontology by two contexts instead of describing each attribute by one modifier context as in atemporal concepts. In query 1 the user expects the currency always be in EGP. While in source 2 (

airline_temporal table) the currency is in EUR if departing date <= "01/08/2012" and changed from EUR to AED when departing date > "01/08/2012".In Source 2 we link the price attribute with C_Before context if the value of Departing_date attribute <= "01/08/2012"and with C_After context if the value of Departing_date attribute >"01/08/2012"

In C_Before context: modifier currency has a value of "EUR"

In C_After context: modifier currency has a value of "AED"

Reconciling Temporal Ontological Conflicts of Query2

In temporal ontological assumptions the ontological concept represented by an attribute may change over time. In airline_temporal table of source 2, Price on and before "01/08/2012" includes taxes, afterwards it excludes taxes .In figure 7 we have two different assumptions (interpretation) for Price (including and excluding taxes). According to our proposed approach we create a more general concept in the ontology that includes both variations as a special cases based on the value of the modifier limit.

In T_Inclusion context: Price (+tax)

In T_Exclusion context: Price (nominal).

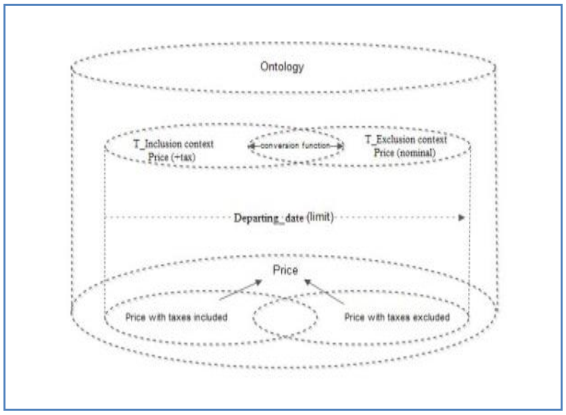
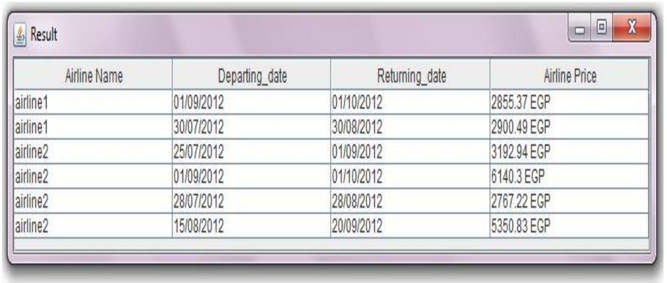


Fig.15. Temporal Ontological Representation

SCR query engine able to handle terms in the ontology that can be expressed based on explicitly define contexts of other terms or adjusting the value of one context to another after assigning the required equations with each context (reconciling equational semantic conflicts).

In source 2 there is an equational conflict between T_Inclusion context and T_Exclusion context so we use the conversion functions associated with each context in the mediated queries in order to reconciling such a conflict.

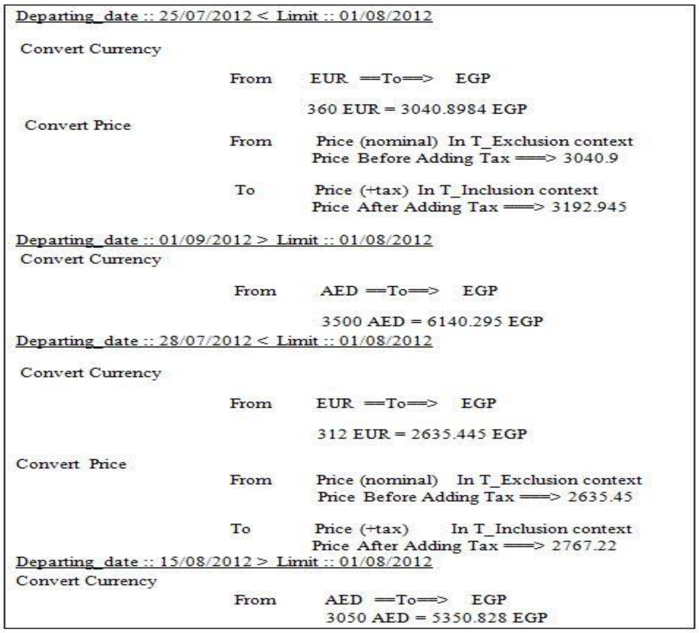
- To reconcile the previous temporal semantic conflicts the converter checks the value of modifier limit in the ontology and the corresponding currency and price modifiers values in order to compare them with the user required context.
- In C_Before context : modifier currency has a value of "EUR"
- In C_After context: modifier currency has a value of "AED"
- In T_Inclusion context : Price (+tax)
- In T_Exclusion context: Price (nominal).



Airline Name	Departing_date	Returning_date	Airline Price
airline1	01/09/2012	01/10/2012	2855.37 EGP
airline1	30/07/2012	30/08/2012	2900.49 EGP
airline2	25/07/2012	01/09/2012	3192.94 EGP
airline2	01/09/2012	01/10/2012	6140.3 EGP
airline2	28/07/2012	28/08/2012	2767.22 EGP
airline2	15/08/2012	20/09/2012	5350.83 EGP

Fig.16. Query1 Final Results

Figure 17 shows trace of reconciling the detected temporal semantic conflicts in Airline2 as follow.



```
Departing_date :: 25/07/2012 < Limit :: 01/08/2012
 Convert Currency
                          From     EUR  ==To==>   EGP
                                   360 EUR = 3040.8984 EGP
 Convert Price
                          From      Price (nominal)  In T_Exclusion context
                                   Price  Before Adding Tax ===> 3040.9
                           To       Price  (+tax)  In T_Inclusion context
                                   Price  After Adding Tax ===> 3192.945
Departing_date :: 01/09/2012 > Limit :: 01/08/2012
 Convert Currency
                          From       AED ==To==>   EGP
                                     3500 AED = 6140.295 EGP
Departing_date :: 28/07/2012 < Limit :: 01/08/2012
 Convert Currency
                          From      EUR  ==To==>   EGP
                                    312 EUR = 2635.445 EGP
Convert  Price
                          From      Price (nominal)   In T_Exclusion context
                                    Price  Before Adding Tax ===> 2635.45
                          To        Price  (+tax)      In T_Inclusion context
                                   Price  After Adding Tax ===> 2767.22
Departing_date :: 15/08/2012 > Limit :: 01/08/2012
Convert Currency
                          From       AED  ==To==>   EGP
                                     3050 AED = 5350.828 EGP
```

Fig.17. Trace of reconciling temporal semantic conflicts in Airlin2

V. Conclusion

We developed an ontology-based approach, in which all data semantics explicitly described in the knowledge representation phase and automatically taken into account by Interpretation Mediation Services phase, conflicts detected and resolved automatically at the query runtime. Unlike the traditional approaches there is no need for any changes for the sources involved in the integration process even if these sources with time-varying semantics, each source should only explicitly records its semantic assumptions in addition to a small number of conversion functions, which are used by the converter for automatically make required conversions. The SCR preserve local autonomy for each data source to change and maintain independently. Data sources still independent from the integration process that is mean we can retrieve up to date data and smoothly update the data in each data source without affecting the integration process. The SCR framework provides a systematic methodology for explicitly describing assumptions through the knowledge representation phase. Changes in sources contexts can be accommodated by modifying annotation values without affecting both the global ontology and the conversion functions (no hand-code needs to be maintained).The proposed SCR framework assumes that the underlying information sources are structured data, testing our proposed approach for detecting and reconciling semantic conflicts using semi-structured data such as web pages is required to identify new challenging issues. We have demonstrated the capability of the SCR framework using simple illustrative example. It is interesting to ensure interoperability and knowledge-based information sharing in real world environments of fertile fields like bioinformatics, digital libraries and GIS systems in order to test the feasibility of our approach.

REFERENCES

[1] Firat, A.(2003). Information Integration Using Contextual Knowledge and Ontology Merging. Ph.D. Thesis. Massachusetts Institute of Technology.

[2] Madnick S., Gannon T., Zhu, H., Siegel M., Moulton A., Sabbouh M.,(2009):" Framework for the Analysis of the Adaptability, Extensibility, and Scalability of Semantic Information Integration and the Context Mediation Approach", IT.

[3] Madnick, S.E., & Zhu, H. (2006) .Improving Data Quality with Effective Use of Data Semantics. *Data and Knowledge Engineering, 59*(2), 460-475.

[4] Noy, N. F. and Musen, M. A., PROMPT: Algorithm and tool for automated ontology merging and alignment, National Conference on Artificial Intelligence - AAAI , pp. 450-455, 2000.

[5] Nyulas, M. Connor, S. Tu, DataMaster_a plug in for Relational Databases into protégé, Stanford University, 10th international protégé conference, 2007.

[6] Ram ,S., Park, J., Semantic Conflict Resolution Ontology (SCROL): A Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts, IEEE Transactions on Knowledge and Data Engineering, v.16 n.2 , p.189-202, 2004.

[7] Rosenthal, A., Seligman, L. and Renner, S. From Semantic Integration to Semantics Management: Case Studies and a Way Forward, ACM SIGMOD Record, 33(4), 44-50, 2004.

[8] Sultan, T. I., Nasr, M. M. , Khedr, A. E., & Ismail , W., S. (2013) " Semantic Conflicts Reconciliation (SCR): A Framework for Detecting and Reconciling Data-Level Semantic Conflicts" , International Journal of Engineering Research and Applications (IJERA),ISSN: 2248-9622 , Volume 3, Issue: 1, pp.766-773, India.

[9] Zhu, H., & Madnick, S. E. (2004) "Context Interchange as a Scalable Solution to Interoperating Amongst Heterogeneous Dynamic Service", *3rd Workshop on E-Business*. Washington, D.C., 150-161.

[10] Zhu, H., (2005): Effective Information Integration and Reutilization: Solutions to Deficiency and Legal Uncertainty, PhD Thesis, Massachusetts Institute of Technology Cambridge, MA, USA.

[11] Zhu, H., Madnick, S., Reconciliation of temporal semantic heterogeneity in evolving information systems", ESD-WP-2009-03, Massachusetts Institute of Technology Cambridge, MA, USA, 2009

An Intelligent mutli-object retrieval system for historical mosaics

Wafa Maghrebi, Anis B. Ammar, Adel M. Alimi
Electrical and Computer Engineering Dept
ENIS, University of Sfax
Sfax, Tunisia

Mohamed A. Khabou
Electrical and Computer Engineering Dept
University of West Florida
Pensacola, USA

Abstract*—In this work we present a Mosaics Intelligent Retrieval System (MIRS) for digital museums. The objective of this work is to attain a semantic interpretation of images of historical mosaics. We use the fuzzy logic techniques and semantic similarity measure to extract knowledge from the images for multi-object indexing. The extracted knowledge provides the users (experts and laypersons) with an intuitive way to describe and to query the images in the database. Our contribution in this paper is firstly, to define semantic fuzzy linguistic terms to encode the object position and the inter-objects spatial relationships in the mosaic image. Secondly, to present a fuzzy color quantization approach using the human perceptual HSV color space and finally, to classify semantically the mosaics images using a semantic similarity measure. The automatically extracted knowledge are collected and traduced into XML language to create mosaics metadata. This system uses a simple Graphic User Interface (GUI) in natural language and applies the classification approach both on the mosaics images database and on user queries, to limit images classes in the retrieval process. MIRS is tested on images from the exceptional Tunisian collection of complex mosaics. Experimental results are based on queries of various complexities which yielded a system's recall and precision rates of 86.6% and 87.1%, respectively, while the classification approach gives an average success rate evaluated to 76%.

Keywords*—*retrieval; mosaics; metadata; classification; multi-objects

I. Introduction

Nowadays many visual information retrieval systems with different complexities and recall capabilities were developed, tested and even made available online. Content Based Image Retrieval (CBIR) approach has been studied and explored for decades [1-5]. However, such systems usually use only a restricted set of low-level features such as color, texture and shape. The features are often computed globally such as in QBIC system [1] or locally such as the FourEyes system [2].

The QBIC system uses texture and color as global features to index images. The global features have some limitations in modeling perceptual aspects of shapes and usually perform poorly in the computation of similarity with partially occluded shapes.

The FourEyes system uses the local features to index an input image; an image is first divided into small and equal square parts then, shape, texture and other local features are extracted from these squares or regions. These local features are then used to index the whole image.

Recently, various museums are constructing digital archives consisting of high-resolution images of paintings and artifacts to preserve the original copies and to make them available to a wider audience via the Internet. For example, the Hermitage Museum of Amsterdam and its partner IBM use a browsing and retrieval system to make images of its collection available online [6]. Vuupijl et al. [7], based on statistic study, demonstrate that 72% of users interest to objects in the images. So, many searches are emerged which give a relevance to the semantic object in the image. For example, Shomaker et al. [8] propose the Vindx system, which uses a cooperative annotation object shape accompanied with a semantic textual classification to index digitized collection of the National Gallery of the Netherlands (the Rijksmuseum[9]).

The Vindx database consists of images of complex paintings from the 17th century containing multiple objects. Some efforts were invested in the improvement of the Vindx system. For example, Broek et al. [10] presented the C-BAR system (Content Based Art Retrieval) to describe the paintings of the Rijksmuseum by the CBIR approach. Broek et al. [11] continue to improve the performance of the system and its user interface. The system developed by Berretti et al. [12,13] indexes objects in an input image based on their shape. Chang et al.[14] developed an XML-based document browsing and retrieval system for the digital museum of Korean porcelain. The high-resolution images consist of various porcelain artifacts photographed on a uniform background.

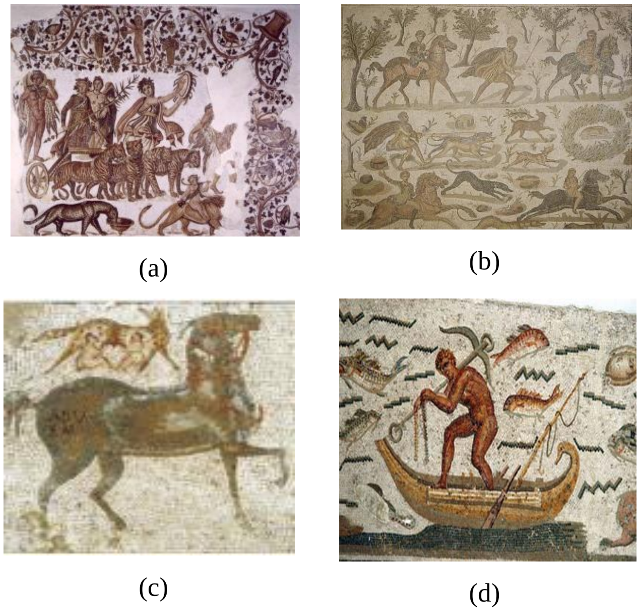
Recently, some researchers have tackled the problem of indexing and cataloging images of mosaics with all the challenges such image particularities present [15-17]. M'hedhbi et al. [16] use a CBIR approach to retrieve mosaics based on shape descriptors. The aim is to dedicate a recognition system to archaeologists. In reference [17] Maghrebi et al., present a retrieval system of Roman mosaics images using drawing queries. They use a robust MPEG7 of low level shape descriptors to index objects. Those retrieval systems [1, 2, 7, 8, 10-17] use, mainly, low level features and don't give the possibility to specify multiple objects in the user query. Nevertheless, a few multi-objects based image indexing and retrieval systems have been developed [18-20]. They have the ability to specify spatial relationships between objects in an image and present to expert users a complex drawing query design.

In this paper we present MIRS, a Mosaics Intelligent indexing and Retrieval System. Our purpose is to extract knowledge from complex mosaics images for multi-object indexing and retrieval. This system has been designed to be user-friendly and to simplify the query process as much as possible.

The paper is organized as follows: In section 2 we start by describing the mosaics database and its particularities. Section 3 discuses the general system architecture. Section 4 details our approach to define mosaics metadata. In section 5 we present the classification approach. Section 6 details the retrieval process. Section 7 is dedicated to the experimentations and results and finally, section 8 concludes the paper.

II. MOSAICS DATABASE

Tunisian museums (e.g. Bardo, El-Jem, Enfidha, Sousse, Sfax) house a huge and exceptional collection of mosaics of great historical value. Some of these beautiful mosaics date back to 420 BC and depict various artistic themes. The mosaics have very rich and complex content consisting of many objects of different shapes, colors, sizes, and textures which make the automatic extraction of meaningful objects from an image very challenging if not impossible. These treasures were carefully photographed and catalogued not only to make them available to researchers, but also to limit direct handling of these fragile articles. Fig.1 shows samples of Tunisian mosaics in our database.



(a) (b)

(c) (d)

Fig. 1. Samples of Tunisian historical mosaics

Mosaic is the art of creating images through assembling small pieces (or tesserae) of colored natural marble. By its nature, this creation technique can cause natural color variation in supposedly uniform regions within the mosaic.

Our database is composed of 200 mosaic images which have been filtered with a Gaussian low-pass filter (σ = 1) to reduce noise caused by the inherent structure of mosaics. We also filtered the images using a 3x3 median filter to reduce the brisk intensity variations within the images.

III. GENERAL ARCHITECTURE OF MIRS SYSTEM

In the proposed system, we use low level features like color and shape descriptors, but we also use high level features such as objects' position, objects spatial relationships. In addition, we propose a fuzzy color quantization approach in HSV color space and a mosaics semantic classification. Our goal is to transform the hard and complex multi-object queries by content to simple textual queries.

As shown in Fig.2, using the objects database the system extracts crisp features dealing with the object region, area, centroid and the smoothed object boundary. These crisp features are used to extract fuzzy linguistic terms to represent the object's color in the perceptual HSV color space, its position in the mosaic, and its spatial relationships with other objects.

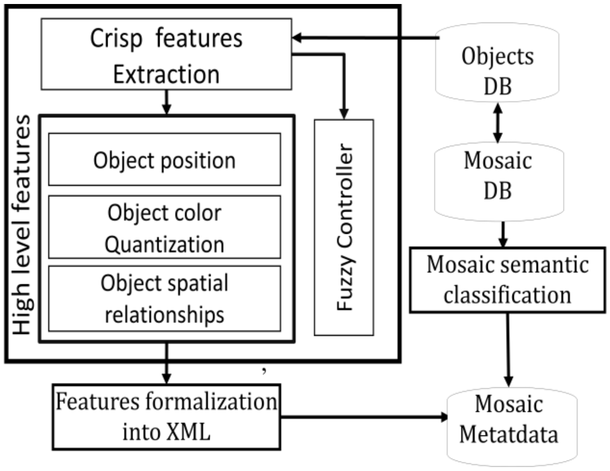

Crisp features Extraction
Objects DB
High level features
Object position
Object color Quantization
Object spatial relationships
Fuzzy Controller
Mosaic DB
Mosaic semantic classification
Features formalization into XML
Mosaic Metatdata

Fig. 2. MIRS: mosaic metadata definition

Our purpose is to attain the mosaic semantic interpretation and to offer to user the ability to handle query that imagines in natural language such as "*man between horse and dog"* or "*poet at the middle near to muse"* or "*Brown or dark brown horse very far from dog*". Both sets of features are formalized into XML language to create the mosaic metadata.

IV. MOSAICS IMAGES METADATA

We apply the fuzzy logic techniques to define the semantic features representing the object position, spatial relationships and color. The semantic linguistic fuzzy features are collected and formalized into XML language to describe mosaics and to facilitate information exchange. The XML language has the capability to self-description, intuitive readable format, simplicity, extensibility since it gives the possibility to add features to description schemas. We detail in the following our approach to define automatically the mosaic semantic features.

A. Fuzzy object position and spatial relationships

The idea is to divide the image into 3x3 regions to define a set of three vertical positions and set of three horizontal ones. We affect to each set fuzzy linguistic terms to encode the object position within the mosaic as a set of membership values to the fuzzy sets *Left, Right, Middle, Up and Down.* The centroid of the object is used in these calculations. For example, the membership to the sets *Left*, *Right,* and *Middle* are calculated based on the horizontal location of the object's centroid within the mosaic using the fuzzy membership functions shown in Fig.3.

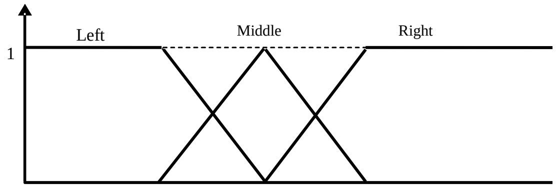


Fig. 3. Membership functions to fuzzy sets Left, Middle, and Right

We enable the users (historians or layperson) not only to specify the objects they are looking for in a mosaic, but also, specify the object position and spatial relationships with the other objects in the image, to enhance the relevance of the retrieved images.

Our approach uses fuzzy logic techniques to extract the spatial relationships between objects. Our aim is to define semantic linguistic terms that precise twelve fuzzy spatial relationships:

Over, Under, On, Beside_of, Very_near, Near_to, Far_from, Very_far_from, At_left_of, At_right_of, between, and *Surrounded_by.* To determine these spatial relationships we have used three steps:

- Definition of the memberships degrees to the fuzzy functions *Beside_of, Near_to, Very_near, Far_from, and Very_far_from.* This step is based on the computation of the relative Euclidian distance between the objects centroids. Using a training image database, we have defined the fuzzy membership function of this subset of spatial relationships as shown in Fig.4

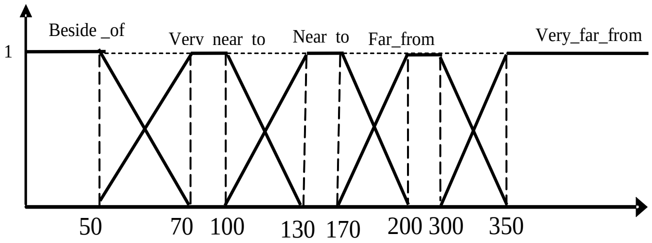


Fig. 4. Fuzzy membership functions to some of the spatial relationships determined based on distances between objects' centroids

- Characterization of spatial relationships *Over*, *Under*, *At_left_of, At_right_of, between*, and *Surrounded_by.* We predefine models for each spatial relationship to describe fuzzy positions neighboring objects. Consequently, we have defined sixty fuzzy rules. Fig.5 below shows some samples of the objects neighboring models that are assimilated to "*Between*" spatial relationships. The evaluation of the fuzzy rules is performed by fuzzy sets operations. We have used *min* and *max* operations for respectively the "*AND*" and "*OR*" operators. After evaluating the result of each rule and in the inference process we retain spatial relationships of activation degree greater than or equal to 0.5.

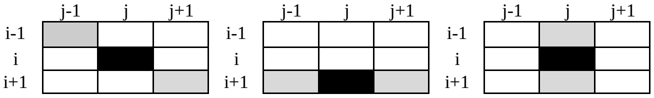


Fig. 5. Samples of "Between" spatial relationships predefined models

- To define the "*On*" spatial relationship, we have applied the Wang similarity measure [21] between the two objects fuzzy positions (i.e. vertical and horizontal). Let $\mu_1(x_i)$ and $\mu_2(x_i)$ with i∈[1,6] vertically and horizontally fuzzy sets of respectively object1 and object2. The Wang similarity measure is defined as[21]:

$$W = \frac{1}{N}\sum_{i=1}^{N}\frac{\min(\mu1(x_i),\mu2(x_i))}{\max(\mu1(x_i),\mu2(x_i))} \quad (1)$$

The proposition "*object1 ON object2"* is true if : i) the objects fuzzy positions similarity greater or equal to threshold t1, ii) the intersection region area between the two object is more than an experimentally predefined threshold t2 and iii) the satisfaction of the following condition : $(y_{2max} > y_{1max})$ with $(x_{1min},x_{1max},y_{1min},y_{1max})$ are the coordinates of object1 region and $(x_{2min},x_{2max},y_{2min},y_{2max})$ the coordinates of object2 region. Table 1 shows samples of fuzzy objects position and spatial relationships formalized into XML language.

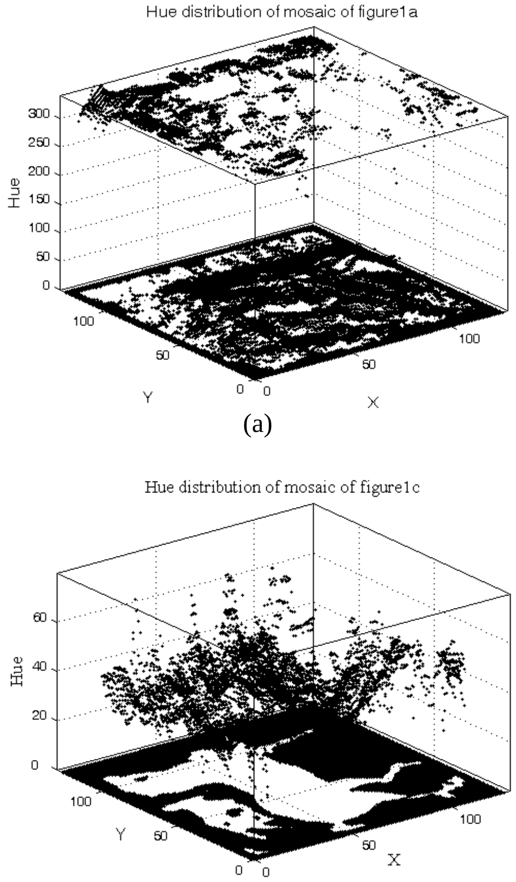
TABLE I. SAMPLES OF FUZZY OBJECTS SPATIAL RELATIONSHIPS FORMALIZED INTO XML LANGUAGE.

Image	Relation
Cleo Virgil Melpomene	<Relation id="41"> < Primary>Virgil </Primary> <relation>**Between**</relation> <second>Cleo</second> <second>Melpomene</second> </Relation> <degree>**1.0**</degree>
	< Relation id ="1"> < Primary > hunter</ Primary > < relation >**On**</ relation > < second > horse </ second > </ Relation > <degree>**1.0**</degree>

B. *Color fuzzy quantization*

Mosaics are a harmony of marble tesserae. So, mosaics are of natural colors. To extract color descriptor, we use the HSV color space (instead of RGB space). The HSV color space is more intuitive and closer to the human perception than the RGB space.

A color is represented in the HSV space by its hue (H) with values between 0 and 360, saturation (S) and value (V) with values between 0 and 100. The value of the hue indicates the nature of the color (e.g. red, blue, etc.), while saturation and value indicate the richness and brightness of the color. For example, Fig.6 shows the hue variations of the sample mosaics in Fig.1 (a and c). Notice that the hue values of the first mosaic (Fig.6a) are generally in the range of [0, 20] or [300, 340], and those of the second mosaic (Fig.6b) are in [0, 60] range. These intervals include variations of reds, oranges and yellows in the hue spectrum. Based on the hue component value of an object, we compute the object's membership values in twelve linguistic fuzzy sets: *red, dark_orange, orange, light_orange, yellow, light_green, green, cyan, light_blue, blue, purpule,* and *pink*. The membership functions of these twelve fuzzy sets are shown in Fig.7.



(a)

(b)

Fig. 6. Hue variations of the two mosaics in Fig.1

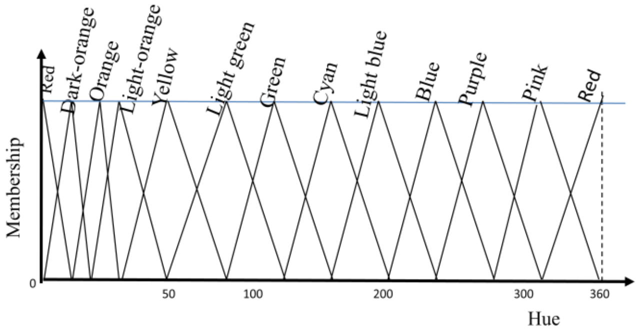
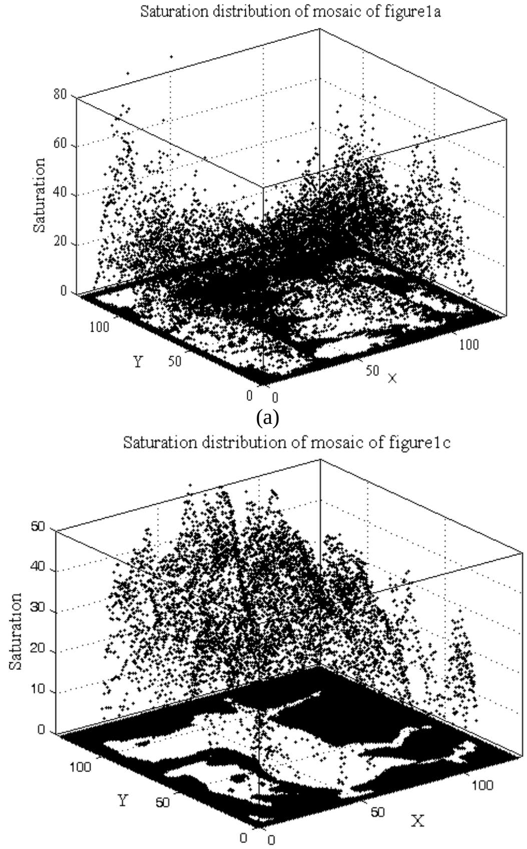


Fig. 7. Hue fuzzy sets

The brightness/darkness of an image is determined by its S and V values in the HSV color space. Fig.8 shows the saturation values S of the two sample mosaics shown in Fig.1 (a and c). In the first sample, the saturation values are mainly between 0 and 40 but can go as high as 80, while in the second sample they are mainly between 0 and 50. Based on these values we create four fuzzy linguistic terms to describe the saturation of an object in the mosaic: *gray, almost_gray, medium* and *clear*.



(a)

(b)

Fig. 8. Saturation variation of the two mosaics in Fig.1

The investigation of the value component V of two samples of mosaics presented in Fig.9 reveals that V ranges between 0 and 70. So, we choose to decompose the V component into four fuzzy linguistic terms: *very_dark, dark, medium* and *light*. Based on the H, S, and V membership degrees of an object, we created a set of 12x4x4=192 rules that define what a human would perceive as the dominant color of an object. Moreover, let's consider the following sample: if H=15, S=100 and V=40. Our fuzzy controller find *dark_brown* color which is presented with fuzzy sets as hue=*dark_orange*, saturation=*clear* and value =*dark*. In addition, for every hue values with *gray* in saturation and *light* in value the color returned is *white*. We find the *black* color if value V is *very_dark* and saturation S is *gray*.

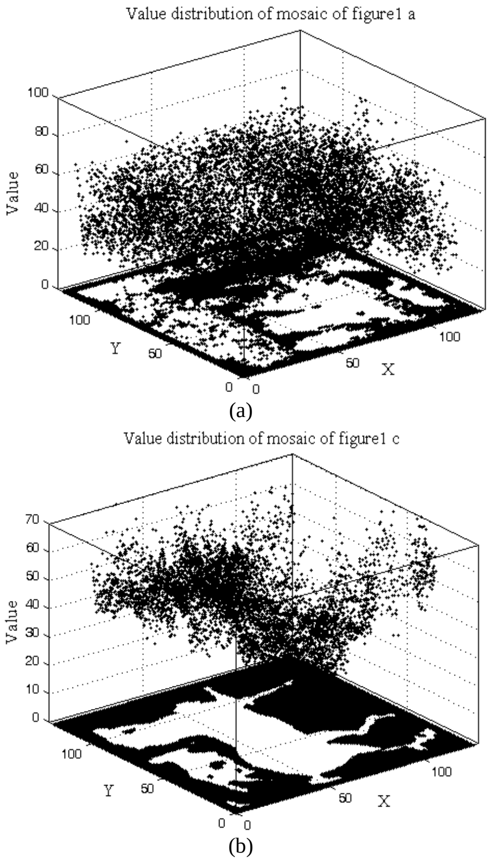
We present in following samples of some predefined fuzzy rules to find equivalent semantic linguistic colors.

If (Hue="*dark_orange*" and saturation="*medium*" and Value="*light*") then color ="*dark_orange*"

If (Hue ="*dark_orange*" and saturation= "*clear*" and Value = "*medium*") then color = "*Honey*"

If (Hue =“*green*” and saturation= “*medium*” and Value = “*dark*”) then color = “*Dark_green*”

If (Hue =“*red*” and saturation= “*almost_gray*” and Value = “*medium* then color = “*Dark_pink*”



Value distribution of mosaic of figure1 a

(a)

Value distribution of mosaic of figure1 c

(b)

Fig. 9. Example of value extracted from two mosaics of Fig.1

The evaluation of the fuzzy rules is performed by the fuzzy sets operations. We have used *min* and *max* operations for respectively the “*AND*” and “*OR*” operators. After evaluating the result of each rule and in the inference process we have used the maximum algorithm as accumulation method. Our approach consists to compute the fuzzy color for each object pixel, determine the object colors histogram and save the five most dominant colors that represent more than 80% of object area. Table 2 shows examples of returned objects colors based on predefined fuzzy rules.

V. Mosaic Semantic Classification

Historians who usually study the mosaics classify them into religion, cultural, social, economic and natural semantic classes. In the religious scenes we find many gods or goddesses such as “Dionysus” the wine god, “Vulcain” the god of fire and “Oceanus” the sea god. Moreover, in these scenes we can find the wild animals such as lions or tigers and/or the imaginary creatures like seahorses, sea panther or centaurs[1].

TABLE II. Examples of Formalized Fuzzy Color in XML

<area>0.85953444</area> <deg_pert>0.990196</deg_pert> <Color>white</Color> <area>0.10990589</area> <deg_pert>0.89215684</deg_pert> <Color>dark_gray_brown</Color> <area>0.00846954</area> <deg_pert>0.7058823</deg_pert> <Color>very_clear_blue</Color> <area>0.008271421</area> <deg_pert>0.7745099</deg_pert> <Color>dark_clear_blue</Color> <area>0.004606241</area> <deg_pert>0.6960784</deg_pert> <Color>dark_blue</Color>	<area>0.42255497</area> <deg_pert>1.0</deg_pert> <Color>black</Color> <area>0.2494867</area> <deg_pert>0.9803922</deg_pert> <Color>dark_gray_brown</Color> <area>0.12088738</area> <deg_pert>0.9526627</deg_pert> <Color>very_dark_gray</Color> <area>0.05623717</area> <deg_pert>0.9375</deg_pert> <Color>very_dark_brown</Color> <area>0.045971256</area> <deg_pert>0.9499998</deg_pert> <Color>dark_brown</Color>

We decided that each mosaic contains god names or imaginary creatures, can be classified as religious scenes. Others mosaics are classified as economic scenes while they present the Roman economic activities such as hunting scenes, olive harvest, or fishing scenes. In some mosaic we find Roman leisure. So, we find circus with many games and celebrate fighting force scenes. In these scenes prisoners or gladiators must realize a fighting with dangerous wild animals (e.g. tiger and bear). Other mosaics show the importance that gives Roman to the literature, in this context we find mosaics which describe the famous musicians (e.g. Orphee) or the Muses who protect cultural activities such as music, songs or poetry. These samples of mosaics can be classified as cultural ones. Mosaics showing the worth of savage animals in the Roman period can be classified as natural scene. To classify the mosaics into these five classes we use semantic concepts and we define for each concept a set of keywords. For example “Dionysus” is a keyword for the “religion” concept. We apply an extensional measure which uses the instances of the concept or the term occurrences that denote the concept in the corpus. This measure is introduced by Sanderson & Croft [22]. They use the probability *p(C)* to have the concept (C) in a given corpus:

$$\mathrm{P(C)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}\in\mathrm{words(C)}}^{\mathrm{N-1}}\frac{\mathrm{Count(n)}}{\mathrm{nbc(n)}} \qquad (2)$$

Where *Count(n)* measures the occurrence number of each concept terms in the corpus, and *nbc (n)* is equal to the number of concepts that the term *n* is a label. This probability measure *p(c)* takes into consideration that one term can be found in one or more concepts. For example the term “*lion*” can be found with “*Dionysus*” in a religion scene, or in wildness scene.

We calculate the *Ψ(c)* introduced by Resnik [23] as followed:

ψ□C)= - log(p(C)) □□□

[1] Centaurs: imaginary creatures from the Greek mythology that presents a head and torso of man and body of horse

The relevant concept *C* of the predefined mosaic is the concept that verifies the following condition:

$$\text{Pertinence (C)= min } \psi(\text{C}) \qquad (4)$$

Where C represents the class number and takes a value between 1 and 5. The application of this approach, on the mosaics metadata, gives in some cases an equal pertinence values between two or more classes. Consequently, it returns more than one class. So, we take all the relevant classes that verify the formula 4. Fig.10 shows two samples of automatic mosaics semantic classification.

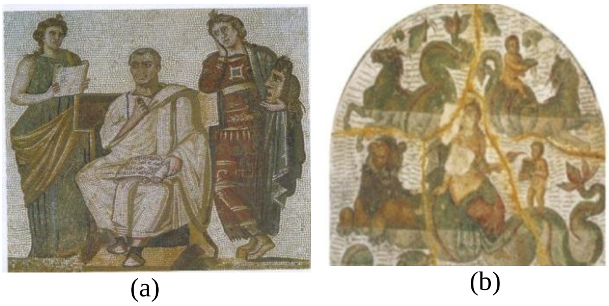

(a) (b)

Fig. 10. Sample of image classification a) Cultural class. The poet Virgil is between the Muses Clio and Melpomene. b) Religion class. Venus, the goddess of love and beauty, is on sealion. The image contains also a seahorse and seapanther.

VI. Query and Retrieval Process

We enable the users not only to specify the objects they are looking for in a mosaic, but also, to specify other semantic descriptors like the object position, semantic color and/or spatial relationships between these objects. Based on the objects described by the user, the system classifies this query using the classification approach explained in section 5 and returns the relevant class(s).

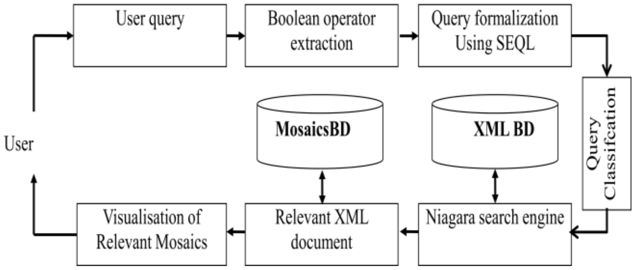
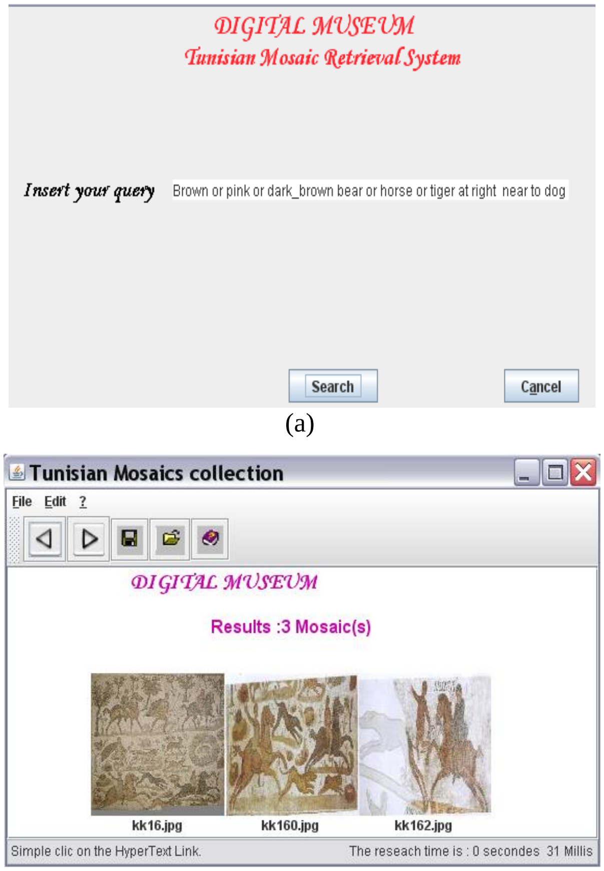

User query
Boolean operator extraction
Query formalization Using SEQL
Query Classifcation
MosaicsBD
XML BD
User
Visualisation of Relevant Mosaics
Relevant XML document
Niagara search engine

Fig. 11. Retrieval process

The architecture of the retrieval system is summarized in Fig.11. The user enters a textual query describing: object(s), object position, object color(s), and object's position relative to other object(s). The system processes the query by filtering the non useful word (e.g. the, a, to, of), parsing Boolean operators (e.g. or, and, not), formalizing the user query into the SEQL (Search Engine Query Language) and, using the integrated Niagara search engine [24,25], returns relevant XML documents and consequently mosaics images corresponding.

The textual query is introduced in the natural language. The retrieval process traits this query automatically and recognize the object(s) textual description, position, color(s) and spatial relationships. The system can be queried with one or more objects and one or more colors using Boolean operators. Fig.12a shows a user looking for mosaics containing "brown or pink or dark brown bear or horse or tiger at right near a dog", and Fig.12b shows the returned results.


DIGITAL MUSEUM
Tunisian Mosaic Retrieval System

Insert your query: Brown or pink or dark_brown bear or horse or tiger at right near to dog

(a)

Tunisian Mosaics collection

DIGITAL MUSEUM

Results :3 Mosaic(s)

kk16.jpg kk160.jpg kk162.jpg

Simple clic on the HyperText Link. The reseach time is : 0 secondes 31 Millis

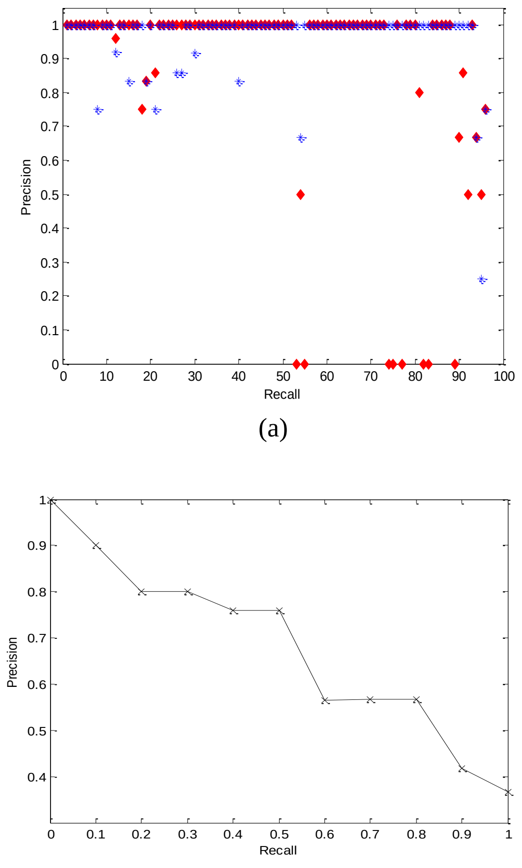
(b)

Fig. 12. The system's query GUI showing: a) a sample query of query in which we search for "brown or pink or dark brown bear or horse or tiger at right near a dog"

VII. Experimental Results

In this study, we use 100 queries to test the system. By these queries and using 1050 objects of our XML metadata, we test the system's recall and precision performance and its ability to accurately handle the semantic queries. The overall system is implemented in Java, using client/server architecture and threads.

As shown in Fig.13a, the precision and recall performance of MIRS retrieval system varies depending on the query, it takes anywhere between 70 and 800 *ms* to process a given query, depending on its complexity. However, we were able to achieve an overall recall and precision rates of 86.6% and 87.1% respectively. Fig.13b shows the precision-recall curve for 40 different queries using the extracted knowledge such as the object position, spatial relationship and color. The precision is found to be a decrease function of the recall and reaches a minimum value of 37% when the recall is equal to 1.



(a)

(b)

Fig. 13. Testing queries with their (a) individual recall (square) and precision (asterisk) rates and (b) precision-recall curve.

In our tests we remark that MIRS gives acceptable results but it presents slight failure in some particular cases. For example we can see in Fig.14 that queries where we wanted a particular *object on another one*, the system gives a good result for 80% of tested images and failure for 20%. This is may be due to the use of some experimentally predefined thresholds t1 and t2, with t1 limits the objects intersection regions chosen to be grater or equal to 0.1, while t2 precise the positions fuzzy similarity and chosen to be greater than or equal to 0.9.

The proposed classification approach is also tested; results give a success average degree evaluated to 91.6% and reach 97.1% if relevant classes are more than one. The mutli-classes returned is may be due to mosaics with multi-scenes such as agriculture and social scenes.

In this case, if the probability to find one concept is greater than the probability to find another concept, our approach returns the most pertinent class. Whenever, if probabilities are equal, this can lead to multi-classes as shown in Fig.15.

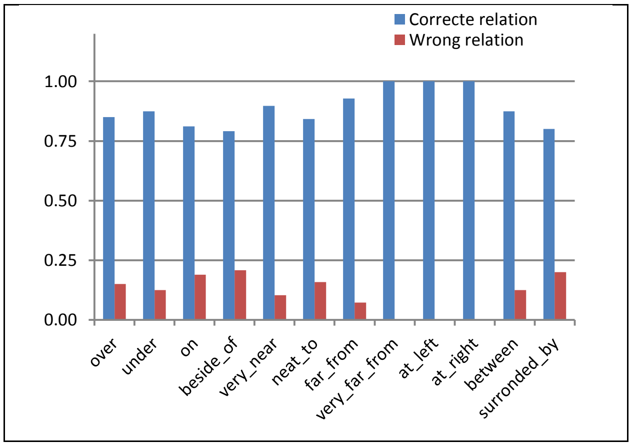


Fig. 14. Statistics of fuzzy object relationships

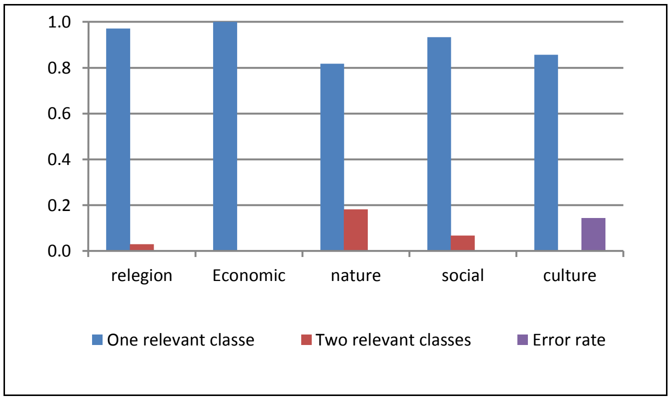


Fig. 15. Statistics on mosaic classification

VIII. Conclusion

In this paper, we present an intelligent system to index and retrieve images of historical mosaics. Our purpose is to transform the hard and complex multi-object queries by content to simple textual queries. The MIRS system proposes a fuzzy quantization color approach in HSV color space. In addition, it extracts semantic linguistic terms using a fuzzy similarity measure and proposes a semantic mosaics metadata classification approach.

The system has an intuitive query and response graphic user interface. The queries are in the natural language and specify the objects textual descriptions, colors, position and/or spatial relationships between them. The system was tested on a database containing 1050 objects, extracted from Tunisian historical mosaics images, by a variety of complex queries. The system was able to achieve respectable recall and precision rates of 86.6% and 87.1%, respectively. The average query processing time was around 400*ms*.

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, “Query by Image and Video Content: The QBIC system,” IEEE Computer, vol. 28, N°9, pp. 23-32, 1995.

[2] T. P. Minka, and R. W. Picard, "Interactive Learning Using a Society of Models," Pattern Recognition, vol. 30 N°4, pp.565-581, 1997.

[3] W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. Pattern Anal. Machine Intell, vol. 22, pp.1349-1380, 2000.

[4] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," Journal of Vision Communication and Image Representation, vol. 10, pp. 39 -62, 1999.

[5] N. Sebe, M. S. Lew, X. Zhou, T.S. Huang, and E.M. Bakker, "The state of the art in image and video retrieval," Int. Conf. on Image and Video Retrieval. Lecture Notes in Computer Science, vol. 2728, pp. 1-8. 2003.

[6] http://www.hermitagemuseum.org Hermitage museum website, 2012.

[7] L. Vuupijl, L. Shomaker, and E. Broek, "Vind(x): Using The User Through Cooperative Annotation," The 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR.8), Canada, pp. 221-225, 2002.

[8] L. Schomaker, L. Vuurpijl, and E. Deleau, "New Use For The Pen: Outline-Based Image Queries," In Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR), Piscataway (NJ), pp. 293-296, 1999.

[9] Rijksmuseum, http://www.rijksmuseum.nl/wetenschap/, 2013.

[10] E. Broek, T. Kok, E. Hoenkamp, Th. E. Schouten, P. J. Petiet, and L. G. Vuurpijl, "Content-Based Art Retrieval (C-BAR)," In Proceedings of the XVIth International Conference of the Association for History and Computing, Amsterdam, pp. 14-17, September 2005.

[11] E.L. Broek, and T. Kok, T. E. Schouten, and L. G. Vuurpijl, "Human-Centered Content-Based Image Retrieval," Human Vision and Electronic Imaging XIII, In Proceedings of SPIE 6806, San Jose, USA, pp. 28-31, January 2008.

[12] S. Berretti, A. Del Bimbo, and P. Pala, "Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing," IEEE Transactions on Multimedia, 2(4), 225 -239, 2000.

[13] S. Berretti, A. Del Bimbo, and P. Pala, "Efficient Matching and Indexing Of Graph Models In Content-Based Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23 N°10, pp. 1089 - 1105, 2001.

[14] J.W. Chang, and Y.J. Kim, "XML Document Retrieval System Supporting Multimedia Web Service for Digital Museum," IEEE international conference on Web services, ICWS, pp. 1001-1007, 2007.

[15] F. Stanco, S. Battiato and G. Gallo, "digital imaging for cultural heritage preservation: analysis, restoration and reconstruction of ancient Artwork," CRC press Taylor and Francis group, 2011.

[16] M. M'hedhbi, R. Mezhoud, S. M'hiri, and F. Ghorbel,. "A new content-based image indexing and retrieval system of mosaic images," In ICT-TA'06, 3rd Int. Conf. on Information and Communication Technologies: from Theory to Applications, pp. 1715–1719, Damascus, Syria, 24-28 April 2006.

[17] W. Maghrebi, A. Borchani, M.A. Khabou, and A.M. Alimi, "A System for Historic Document Image Indexing and Retrieval Based on XML Database Conforming to MPEG7 Standard. Graphics Recognition," Recent Advances and New Opportunities, LNCS vol. 5046, pp. 114-125, 2008.

[18] G. Scott, M. Klaric and C. Shyu, "Modeling multi-object spatial relationships for satellite image database indexing and retrieval," CIVR, LNCS, vol. 3558, 247-256, 2005.

[19] C.S. Li, J.R. Smith, L.D. Bergman, V. Castelli, "Sequential processing for content-based retrieval of composite Objects," proceeding SPIE/IS&T Symposium on electronic Imaging: Science and Technology – Storage 1 Retrieval for Image Video Databases VI, 1998.

[20] Katare, S.K. Mitra and A. Banerjee, "Content based image retrieval system for mutli object image using combined features," Proceedings of the international conference on computing: theory and applications ICCTA, 2007.

[21] W.J. Wang, New similarity measures on fuzzy sets and on elements, Fuzzy sets and systems, vol. 85, pp. 305–309, 1997.

[22] M. Sanderson and W. Croft, "Deriving concept hierarchies from text," In Proceedings of the 22nd International ACM SIGIR Conference, pp. 206–213, 1999.

[23] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its Application to Problems of Ambiguity in Natural Language," Journal of Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.

[24] J. Naughton, D. Dewitt, D. Maier, A. Aboulnaga, J. Chen, L. Galanis, J. Kang, R. Krishnamurthy, Q. Luo, N. Prakash, R. Ramamurthy, J. Shanmugasundaram, F. Tian, K. Tufte, and S. Viglas, "The Niagara Internet Query System," IEEE Data Engineering Bulletin, vol. 24, pp. 27-33, 2001.

[25] NIAGARA Query Engine, *http://www.cs.wisc.edu/niagara/*, 2013.

Optimization Query Process of Mediators Interrogation Based on Combinatorial Storage

L. Cherrat, M. Ezziyyani, M. Essaaidi
University Abdelmalek Essaadi, LaSIT
F.S. Tetuan, Morocco

Abstract—**In the distributed environment where a query involves several heterogeneous sources, communication costs must be taken into consideration. In this paper we describe a query optimization approach using dynamic programming technique for set integrated heterogeneous sources. The objective of the optimization is to minimize the total processing time including load processing, request rewriting and communication costs, to facilitate communication inter-sites and to optimize the time of data transfer from site to others. Moreover, the ability to store data in more than one centre site provides more flexibility in terms of Security/Safety and overload of the network. In contrast to optimizers which are considered a restricted search space, the proposed optimizer searches the closed subsets of sources and independency relationship which may be deep laniary or hierarchical trees. Especially the execution of the queries can start traversal anywhere over any subset and not only from a specific source.**

Keywords—Mediation; Datawarehouse; Optimisation; Classification

I. Introduction

The challenge created by the increase and the diversity of information sources on the web, and by the need of organizations to interoperate database systems not only consists of the need to use tools for integrating data [3][5][9][10] among multiple users and heterogeneous information sources, but also the necessity of these tools to overcome the limitations of current search engines by allowing not only users to ask queries more sophisticated than simple keywords, but also being able to aggregate other elements of answers from different sources to build, in the most optimized possible way by time and space research, the analytical global response to the user query. This need is becoming increasingly relevant for medical information, especially with the existence of a multitude of web sources specific to medicine areas and the trend towards computerization of patient medical records [2].

Since query processing of data integration [1][6] [11][12] requires access to the data from numerous wide distribution sources over network, it is crucial to investigate how to deal with the expensive communication over head and the response time. In this paper, we present an efficient approach for processing distributed sources with the existence of an execution order graph [2]dependency of the integration system. In the first of a given set of sources, the algorithm classifies the integrated sources into non-exclusive groups (local data warehousing), such that the associate operations can be locally processed without data transfer. Local data warehousing offers many benefits: reduced costs, increased flexibility, and simplified data access with greater agility. Indeed, local data warehousing offers power to interrogate several centralized sources, but also the possibility to analyze the data more efficiently and with low cost on any server based on availability and needs. This solution effectively enabling more users to access more and more data with more ease. Thanks to the Distributed Databases Solution, we can migrate critical data on data centers and improve the response time of readjustment and equilibration of the data distribution. In this perspective, the use of the principle of local data warehousing report a very suitable solution for the integration systems [8].

Our goal is to create disjoint subsets of sources with low coupling the maximum possible. The question is: on what criteria we will classify sources into a disjoint data warehouse? To do this, we develop a relevant algorithm for grouping the sources into subsets based on a new classification method that we propose in this paper.

In the remainder of this paper, we start with Section II by introducing our query optimization method based on the sources classification and the used classification techniques. We next, in Section III, present the sources classification principle and the generated algorithms. In Sections IV, we develop a new method for refining the regrouping result in the aim to readjust the subsets generated by our hybrid classification algorithm. We then in section V, study the performance of data transmission on the network during the interrogation of the mediator with the presence of local data warehouses generated by the algorithm that we proposed. Section VI concludes with future agenda.

II. Query Optimization Based on the Sources Classification

In order to optimize the process of querying sources integrated by the mediator [2][4], we proceed to the construction of partitions of a set of homogeneous sources with known distances and similarities between pairs of sources. Both functions are defined by the degree of dissimilarity and similarity [24] between sources based on the structure of the schema and the data recorded in sources. To do this, we use the approaches and methods of partitioning based on the optimization algorithm which allows us to find a lower cost solution for each partition with the consideration of the homogeneity of the sources in the same partition. Generally, each partition founded by the global optimization algorithm cannot meet the basic constraints of the predefined partitioning. The algorithm then, proceeds to the error correction intra-partition. Such a process is called refining process, which consists of the refining of a partition to increase its homogeneity.

Refining algorithms are used to distribute the sources in the partitions satisfied the constraints of homogeneity and distribution and they have two common objectives: (i) find a partition such that the objective functions, distance and similarity, take respectively the minimum and maximum value. (ii) find a partition such that the variance of homogeneity partitioning is respected as much as possible. The difference between partitioning methods vary according to the order of priorities between these two objective functions. In our algorithm, we give more importance to minimize the distance function, when the similarity functions [19] (load distribution), it will have as a primary goal to respect the homogeneity of inner-partitions. In this paper, we propose two new methods for classification using a hybrid combination of the two following classic classification techniques [23] :

a) Hierarchical approach: It is based on the following principle: create a set of partition distributed hierarchically into disjoint groups (ie into partitions with less and less parts). Each new partition is obtained by successive grouping of parts of the partition immediately preceding in the hierarchy. The sets of sources are divided into two groups to form a tree whose top node is represented by the set of sources and the subset element by the two partitions and so on for each subset created.

b) Mobile centers Method: This is an iterative method that consists of calculating the center of gravity for each part of the partition, and to recreate a partition where each part consists of the nearest elements to the center of gravity. The center of gravity is calculated based on the weight of the global schema. The distance between the global schema and a source is calculated based on the similarity function between the source and the global schema. The next section presents our hybrid method for partitioning sources.

III. Sources Classification Rules

The natural solution to this question is to maintain a distributed data warehouse, consisting of multiple local sources adjacent to the collection points, together with a coordinator. In order for such a solution to make sense, we need a technology for the data classification process [7]. We have developed a new algorithm for this task. This algorithm translates a set of sources into distributed distinct subsets and generates distributed data warehouses, with the following rules: (i) each generated data warehouse performing some computation and communicating the query result to the coordinator, and (ii) the coordinator synchronizing the results and (possibly) communicating with the data warehouses. The semantics of the subqueries generated by system ensure that the amount of data that has to be shipped between data warehouses are independent and use the underlying data. The solution allows for a wide variety of optimizations that are easily expressed in the interrogation and thus readily integrated into the query optimizer. The optimization algorithm included in our prototype contributes to the minimization of synchronization traffic and the optimization of the data processing at the local sites. Significant features of the this approach are the ability to perform both distribution and optimization that reduce the data transferred and the number of evaluation rounds.

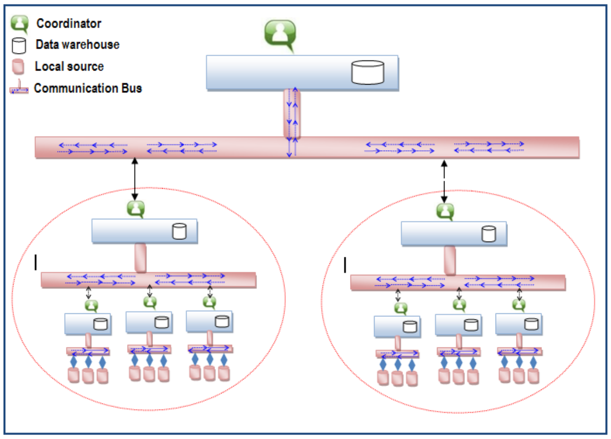


Fig .1. Sources communication

A. *Principle of classification*

The basic idea of this solution is: data in the network is transmitted as a small fragment from set of sources to others, which is obviously a non-redundant way. When data is transferred to another venue, not every datum is involved in connection operation nor useful. Therefore, the data is not involved in the connection, for useless data needs not be transmitted circularly in the network. The basic principle of this optimization strategy is to use the local data warehouse to only transmit the data involved in the connection in the network as far as possible.

The interrogation of the hybrid mediator generally performed in accordance with the created relationships between local data warehouses. Its advantage and deficiency is not considered how to optimize the order of the sub-query to further reduce the network communication costs. But we consider this task it was taken on consideration at the order optimizer process. The solution in this paper is presented according to the deficiency of general algorithm, that is, through the cost estimate to generate the local data warehouses and interrogation process to improve how to further reduce the transfer data cost of sub-query. In this paper we will demonstrate, the results data generated by performing all the sub-queries and generating the final result are regarded as the decisive factor of the creation of the local data warehouses, and the optimization benefits of the order execution.

B. *Source classification algorithms.*

In this section, we propose a new method for the classification of sources based on the principle of the top-down hierarchical method and the mobile centers method. Indeed, this hybrid classification method is based on the knowledge of a distance function and a function of dissimilarity between all pairs of sources of the set integrated by the mediator. In the first, we propose a solution which is based on the principle of top-down hierarchical in the perspectives to improve it with the introduction of the method of mobile centers.

To do this, we define a function that calculates the distance between pairs of sources. However, there is no immediate relationship between distances for all sources of a graph of sources. If the relationship of a distance can be established, it is generally very expensive to implement, especially for non-related graphs.

Therefore, classification methods by graph partitioning are generally impracticable. To some extent, the ascending hierarchical methods could be used without the knowledge of the distance between each source. In this case, they will work nearer to nearer from the known distances between neighboring elements. In this adaptation, each element is a top of the graph, and the distance between neighbors is the cost associated with the edge connecting this top to another top. In fact, such partitioning approaches for graph of sources, are known as the methods of expanding region.

C. Used Functions for the classification of sources

In this section, we are interested to grouping the sources into subsets such that the sources of the same set react similarly to changes of user queries. These problems are often treated with automatic classification methods [20][21] to identify groups of data sources with a homogeneous behavior or quasi-homogeneous to generate a result for the same query to form groups of homogeneous sources, i.e. groups of sources such as sources are as similar as possible within a group (compactness criterion), and the groups are as dissimilar of the similarity and the dissimilarity is based on the set of the following variables :

as possible (criterion of dissimilarity). The measurement The structure of the database schema.

- The nature and number of attributes of entities.
- The size and occurrence of records.
- The inter- entities relationship.
- Results of requests for canonical query (Standard).

1) Distance Function:

- $\mathrm{NbrE}(\mathrm{S_i})$ **:** is the number of entities in the source $\mathrm{S_i}$
- $\mathrm{NbrE}(\mathrm{S_j})$: is the number of entities in the source $\mathrm{S_j}$
- $\mathrm{NbrAtt}\left(\mathrm{E_k^{S_i}}\right)$: is the attributes number of the entity $\mathrm{E_k^{S_i}}$
- $\mathrm{NbrAttId}\left(\mathrm{E_k^{S_i}}, \mathrm{E_l^{S_j}}\right)$: is the number of the identical attributes between the two entities$\mathrm{E_k^{S_i}}$ and $\mathrm{E_l^{S_j}}$.

2) Similarity function and coupling function

To measure the similarity between two sources, we adopt the cosinus rules [26]. With the sets are the sources and elements are the entities. Therefore, we define a similarity function between two sources: it is the report of intergroup rapprochement. The second function is the function allowing to calculate the intra-sources coupling ratio between two sources of the same group. Both functions are based on the weight of the source for each entity. In the next section, we present the data mathematical model used to define these functions.

Let $\mathrm{S} = \mathrm{E_1}, \mathrm{E_2} \dots, \mathrm{E_N}$ a set of entities of the source S. We define the weight of the entity by the number of attributes, the

We define in this section the Distance function [25] between two sources $\mathrm{Distance}\left(\mathrm{S_i}, \mathrm{S_j}\right)$ which is mainly based on the difference between the metadata of the two sources. Indeed, the value of the distance function depends on the number of distinct attributes between all pairs of entities from two sources. The principle of this function is to calculate the distance between two vectors in space. To do this, we assume that each source is a vector whose coordinates are the entities of the source. Thus, the distance between the two sources is the Euclidean Distance between two vectors. We therefore define this function as follows:

$$Distance\left(S_i, S_j\right) = \sqrt[2]{\sum_k^{NbrE(S_i)} \sum_l^{NbrE(S_j)} (DE(E_k^{S_i}, E_l^{j}))^2} \quad 1)$$

With:

$DE(E_k^{S_i}, E_l^{S_j})$: Is the distance between the entity $\mathrm{E_k^{S_i}}$ of the source $\mathrm{S_i}$and $\mathrm{E_l^{S_j}}$ of the source $\mathrm{S_j}$. such as:

$$DE\left(E_k^{S_i}, E_l^{S_j}\right) = \frac{NbrAtt\left(E_k^{S_i}\right) - NbrAttId\left(E_k^{S_i}, E_l^{S_j}\right)}{NbrAtt\left(E_k^{S_i}\right) + NbrAtt\left(E_k^{S_i}\right)} + \frac{NbrAtt\left(E_l^{S_j}\right) - NbrAttId\left(E_k^{S_i}, E_l^{S_j}\right)}{NbrAtt\left(E_k^{S_i}\right) + NbrAtt\left(E_k^{S_i}\right)}$$

$$DE\left(E_k^{S_i}, E_l^{S_j}\right) = \frac{NbrAtt\left(E_k^{S_i}\right) + NbrAtt\left(E_l^{S_j}\right)}{NbrAtt\left(E_k^{S_i}\right) + NbrAtt\left(E_k^{S_i}\right)} + \frac{-2 * NbrAttId\left(E_k^{S_i}, E_l^{S_j}\right)}{NbrAtt\left(E_k^{S_i}\right) + NbrAtt\left(E_k^{S_i}\right)}$$

With:

number of recordsets and the relation with the other entities of the same source.

$$\mathrm{P_E} = \sum_{\mathrm{i}} (\mathrm{Card(E)}) * (\|\mathrm{E}\| - \mathrm{Nb_{FK}(E)}))$$

$\mathrm{Card(E)}$: Number of recordsets of the entity E.

$\|\mathrm{E}\|$: Number of Attributes of the entity E.

$\mathrm{Nb_{FK}(E)}$: Number of external key of the entity E.

$\mathrm{Similitude}(\mathrm{S_i}, \mathrm{S_j})$ is the degree of similarity between two sources Si and Sj. that is to say, the similarity between two sources regarding the schema structure constituting the two sources In this case, the value of the $\mathrm{Similitude}(\mathrm{S_i}, \mathrm{S_j}) \in$ [0,1]. To calculate the similarity between the two sources, we use the Cosinus similarity. Indeed, given two sources $\mathrm{S_i}$ and $\mathrm{S_j}$. The similarity $\mathrm{Cosinus}\left(\theta_{(i,j)}\right)$ is represented by using a scalar product and a grandeur value, which is defined as follows:

$$\mathrm{Similitude}(\mathrm{S_i}, \mathrm{S_j}) = \mathrm{Cosinus}\left(\theta_{(i,j)}\right) = \frac{\mathrm{S_i} \cdot \mathrm{S_j}}{\|\mathrm{S_i}\| \|\mathrm{S_j}\|}$$

$$= \frac{\sum_{k}^{n}(P_{E_i^k}) \times \left(\sum_{l}^{m}(P_{E_i^l}\right)}{\sqrt{m \times \sum_{p}^{n}((P_{E_i^k})^2} \times \sqrt{n \times \sum_{l}^{m}(P_{E_i^l})^2}}$$

The resulting similarity ranges which tend to 0 means exactly that two sources are disjoint. If the value is 1, it means that the two sources are identical. For other values, it indicates the degree of similarity or dissimilarity between the two sources.

Subsequently, we define the coupling function between two sources S_i and S_j, signify the probability of executing a query with the interrogation of the two sources S_i and S_j to generate the result. To do this, we use the Jaccard similarity coefficient [25][26]. The Jaccard coefficient measures the similarity between two sources, it is defined as the ratio of the number of common attributes between the two sources on the number of the union of attributes of two sources:

$$Jaccard(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \quad (1)$$

The Jaccard distance $\mathrm{Jaccard}(S_i, S_j)_\delta$ measure the dissimilarity between sets. It consist simply to subtract the Jaccard coefficient to 1 $(1 - \mathrm{Jaccard}(S_i, S_j))$. Therefore, the coupling function between two sources Sj and Si is a function that gives the degree of similarity between two sources (inter-group) belonging to the same group. This is the relation between the similarity and the distance between two sources proportionally to the weight of the intersection of the two sources.

$$\mathrm{Coupling}(S_i, S_j) = P_{S_i \cap S_j} \times \frac{\mathrm{Jaccard}(S_i, S_j)}{\mathrm{Jaccard}(S_i, S_j)_\delta}$$

$$\mathrm{Coupling}(S_i, S_j) = P_{S_i \cap S_j} \times \frac{\mathrm{Jaccard}(S_i, S_j)}{1 - \mathrm{Jaccard}(S_i, S_j)}$$

$$\mathrm{Coupling}(S_i, S_j) = P_{S_i \cap S_j} \times \frac{\mathrm{Jaccard}(S_i, S_j)}{\frac{|S_i \cup S_j| - |S_i \cap S_j|}{|S_i \cup S_j|}}$$

$$\mathrm{Coupling}(S_i, S_j) = P_{S_i \cap S_j} \times |S_i \cup S_j| \times \frac{\mathrm{Jaccard}(S_i, S_j)}{|S_i \cup S_j| - |S_i \cap S_j|}$$

$$\mathrm{Coupling}(S_i, S_j) = P_{S_i \cap S_j} \times |S_i \cup S_j| \times \frac{\mathrm{Jaccard}(S_i, S_j)}{Distance(|S_i \cup S_j|, |S_i \cap S_j|)}$$

Note: The function $\mathrm{Coupling}(S_i, S_j,)$ is used during the process of refining and readjustment groups (section V).

D. First classification method""

This classification method into clusters seeks to find for each source, all other sources such as distances to this source is minimal and the similarity is maximal. To do this, we use the deviation parameter of the distance $\boldsymbol{\varepsilon_{S_i}^{Dis}}$ and the similarity $\boldsymbol{\varepsilon_{S_i}^{Sim}}$ for the sources $\mathbf{S_i}$.

The deviation of distances for a given source $\mathbf{S_i}$ is :

$$\boldsymbol{\varepsilon_{S_i}^{Dis}} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{NbrS - 1}} \sum_{\boldsymbol{j=k}}^{\boldsymbol{NbrS}} \left(\boldsymbol{Distance(S_i, S_j) - Moy_{Dis}(S_i)}\right)^2}$$

$$Moy_{Dis}(S_i) = \frac{1}{NbrS} \sum_{j=k}^{NbrS} \left(Distance(S_i, S_j)\right)$$

With: NbrS : is the total number of sources

The Deviation for similarities for a given source $\mathbf{S_i}$ is:

$$\boldsymbol{\varepsilon_{S_i}^{Sim}} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{NbrS - 1}} \sum_{\boldsymbol{i=k}}^{\boldsymbol{NbrS}} \left(\boldsymbol{Similitude(S_i, S_i) - Moy_{Sim}(S_i)}\right)^2}$$

$$Moy_{Sim}(S_i) = \frac{1}{NbrS} \sum_{j=k}^{NbrS} \left(Similitude(S_i, S_j)\right)$$

With: NbrS : is the total number of sources

We define the group of the source $\mathbf{S_i}$ by the intersection of the two groups $G_{S_i}^{D}$ et $G_{S_i}^{S}$ as follows:

$$\boldsymbol{G_{S_i}^{D} = \{S_j \in S \,/\; Min_{S_i}^{Dis} - \varepsilon_{S_i}^{Dis} < Distance(S_i, S_j) < Min_{S_i}^{Dis} + \varepsilon_{S_i}^{Dis}\}}$$

and

$$\boldsymbol{G_{S_i}^{S} = \{S_j \in S \,/\; Max_{S_i}^{Sim} - \varepsilon_{S_i}^{Sim} < Similitude(S_i, S_j) < Max_{S_i}^{Sim} + \varepsilon_{S_i}^{Sim}\}}$$

thus :

$$\boldsymbol{G_{S_i} = G_{S_i}^{D} \cap G_{S_i}^{S}}$$

To generate different classification groups with a recursive manner, we follow the following steps:

1) *Initialize G by set of sources.*
2) *Wile G $\not\equiv \phi$*
   - Select one source $S_x \in G$, we search group G_{S_x} identical to G_{S_i}. Thus, the group is the union of all groups.
   - $G \Leftarrow G \setminus G_{S_x}$

Below the generic algorithm of the classification method presented in the previous section:

1) Selection of distances and the grouping method.

2) Calculating the distance between all pairs of individuals (matrix).

3) Each individual is considered a cluster.

4) Research of the two clusters to combine (cf. clustering method [14][15][16][22]).

5) Merging of the two clusters and update the distance matrix.

6) *Repeat from steps 4 until you have only one cluster*.

1) Analysis of result.

The problem is due to the global differences in the degrees of belonging of a source to a given set. It may happen that a source has a distribution of entities similar on two sets, but for one of the two, the degree of belonging is always smaller than the other. We can consider that this source stores the same data and that one of the two sources includes the other, or that one source have a cardinality less than the other. However, as the Euclidean distance is based on absolute differences, these two sources are probably distant and therefore classified in different categories. We say that there is a "Size Effect".

We can overcome this problem by generating two new sources by bursting the source in question. But this transformation does not solve all problems. Indeed, if several variables are related to the same underlying phenomenon, they will be correlated between them and provide the same information several times.

To avoid this drawback, this method can be improved by the use of, on one hand, a fixed number of predefined subsets. Each subset has a center of gravity represented by the local schema. On the other hand, by the separate use of the distance function and the similarity (see the following section) that engages the first for defining sets and the second for correcting error of intra-group belonging.

E. The second method " Gravity Center"

This algorithm aims to build a set of disjoint partitions of all the integrated sources. At the beginning of the algorithm, it is necessary to fix a number k of groups and choose an initial partition. The number of the partition can be inspired by a priori knowledge of the application areas integrated by the mediator. In this method, we adopt the rules of the center of gravity based on the sources local schemas (SL) for all predefined sub-domains. This requires prior knowledge of the primary domain integrated by system and the sub-domains its which composed with local schemas. For each sub-domain, we define a local schema to represent the center of gravity $\mathrm{Cg_i}$ for a group around the center. Then, based on both distance and similarity functions presented in Section V3, to seek all sources belonging to this group. To do this, we minimize the distance and we maximize the similarities between the sources and the center of gravity of each group according to the values of the deviations $\varepsilon_{\mathrm{G_i}}^{\mathrm{Dis}}$ and $\varepsilon_{\mathrm{G_i}}^{\mathrm{Sim}}$. The calculation of the latter depends on the number of sources, the number of groups and the average distance from sources to gravity center. The process starts by the generation of the group whose gravity center has the greatest weight while taking all sources into account. Subsequently, the second group will be formed with the inclusion of unaffected sources to the previous groups, and so on. The group's center of gravity $\mathrm{Cg_i}$ is the intersection of the two groups $\mathrm{G_{S_i}^D}$ and $\mathrm{G_{S_i}^S}$ such:

$$G_{g_i}^{Dis} = \{S_j \in S \,/\; \boldsymbol{Dis}(Cg_i, \boldsymbol{S_j}) < \boldsymbol{\varepsilon_{G_i}^{Dis}} \;\; and \;\; \boldsymbol{S_j} \notin G_{g_k}^{Dis}, \curlyvee k < i\}$$

And

$$G_{g_i}^{Sim} = \{\, S_j \in S \,/\; \boldsymbol{\varepsilon_{G_i}^{Sim}} < \boldsymbol{Sim}(Cg_i, \boldsymbol{S_j}) \; and \;\; \boldsymbol{S_j} \notin G_{g_k}^{Sim}, \}$$

We assume that $P_{G_{g_i}^{Dis}} < P_{G_{g_j}^{Dis}}, \curlyvee i < j$ and $P_{G_{g_i}^{Sim}} < P_{G_{g_j}^{Sim}}, \curlyvee i < j$

With $\varepsilon_{\mathrm{G_i}}^{\mathrm{Dis}}$ and $\varepsilon_{\mathrm{G_i}}^{\mathrm{Sim}}$ are the standard deviations of the set S.

Thus : $Gg_i = G_{g_i}^{Dis} \cap G_{g_i}^{Sim}$

The classification algorithm using the method of gravity center is as follows:

1) Initialize $\boldsymbol{S}$ by all sources.

2) Determine all the centers of gravity $\boldsymbol{Cg_i}$ represented by the local schema of application sub-domain (K centres).

3) Calculate the weight of each center of gravity.

4) Sort the K centers of gravity in descending order by weight.

5) For each center of gravity $\boldsymbol{Cg_i}$, calculate the standard deviations of the set $\boldsymbol{S}$: $\boldsymbol{\varepsilon_{G_i}^{Dis}}$ $\boldsymbol{and}$ $\boldsymbol{\varepsilon_{G_i}^{Sim}}$

a. Compute the set$G_{g_i}^{Dis}$.

b. Compute the set$G_{g_i}^{Sim}$.

c. Determine the group $Gg_i = G_{g_i}^{Dis} \cap G_{g_i}^{Sim}$.

d. Initialize $S = S \setminus Gg_i$.

e. If $S = \{\emptyset\}$, $Exit \;\; loop$.

IV. Refining the Regrouping Result

A. Refining principle

The execution of the hybrid classification algorithm that we proposed in the previous section can automatically generate a set of groups (subsets) that respects the basic constraints defined by the objective function of the hybrid classification algorithm, but does not take into account the general context of the application domain. Therefore, two sources of the same subset generated by the algorithm may have a low semantic relationship, but belong to the same subset according to the principle of the gravity center classification algorithm used in our algorithm. Otherwise, two different sources can have a strong semantic relationship between them, but belong to two different subsets. This means that a refining processor is essential for readjustment of subsets generated by our classification algorithm.

This step aims to minimize the cost of data exchange during the execution of subqueries on geographically remote sites. We propose in this section, the refining process with double treatment: Inter-subset and intra-subset. To define this refining process, we describe in the next section a coupling function between two subsets which gives the degree of correlation (and/or isolation) between groups. Generally, we separate between three possible situations:

1) *Isolated Subset*

Isolated subset is a subset without data replication which has very low coupling (NULL) with other subsets generated by the algorithm of classification. In this case, we can ignore the cost of exchanges between the two subsets. Therefore, we do not apply the refining process on this set for a readjustment. So, it is the very high condition of the end refining process.

2) *Low couplet subset with other subsets*

This is a subset with a data replication and low coupling between all other subsets generated by the algorithm. The threshold value of the low coupling is defined during the configuration of quality of service (QoS) parameters of the classification algorithm. In this case also, we can ignore the cost of exchanges between tow subsets. Therefore, we do not apply the refining process on this set for a readjustment. This is the condition acceptably low for the end of the refining process.

3) *Highly (or strong) couplet subset with other subsets*

This is a subset with a data replication and highly coupling between all other subsets generated by the algorithm. In this case, the cost of exchanges between the two subsets may influence the quality of the algorithm. Therefore, we apply the refining process on this set for a readjustment. This is the condition for the continuation of the refining process. In this case, we proceed to the creation of another subset of sources such that the new subset will allow us to minimize the exchange of data between sources during a query process.

B. *Subsets readjustment algorithm*

The basic idea of the subsets readjustment proposed in this paper is to either move the sources of low coupling with other sources of the same group to groups of highly coupling, or to create a new groups.

The transfer or change of sources is based on the criterion of belonging. The criterion for membership of a source to a group depends on the threshold value proposed by the administrator system as a parameter of quality of service as we will define in the next section. For a description of this algorithm, we propose the following data model:

- Threshold (G) : the minimum threshold for the validation of belonging a source to a group G. It is defined as follows:

$$\text{Threshold (G)} = (\text{Max Coupling}(\text{Si}, \text{Sj}) \times \text{Min Distance (Si,Sj)})$$

- We assume that the source S belongs to the group G. We define the belonging degree of S to G and we denoted by DA(S,G), by the proportional ratio of the sum of the similarities of the source S with other sources of the same group and the sums of the distances from the source S to the sources of the other groups.

$$\mathrm{DA}(\mathrm{S},\mathrm{G}) = \sum \mathrm{Coupling}(\mathrm{S},\mathrm{Si}) \times \frac{\sum \mathrm{Similitude}(\mathrm{S},\mathrm{S_i}))}{\sum \mathrm{Distance}(\mathrm{S},\mathrm{S_j}))}$$

With: $S_i \in G$ and $S_j \notin G$.

- Therefore, we define LowCoupling (G) by all sources of low coupling of the group G. This is the set of sources with the value of the degree of belonging validation is less than the threshold of G.

$$\text{LowCoupling (G)} = \{\text{Si} / \text{DA (Si, G)} < \text{threshold(G)}\}$$

So during the adjustment process or the refining of each group G, we begin with the generation of all the sources of low coupling for each not empty group G (LowCoupling (G)) and for each source S_i of this set, we proceed to the following steps:

- If all the belonging degrees of the source to the other groups are less than the belonging validation thresholds, we assign this source Si to a new group.
- If not, the source Si is added to the group of a maximal validation belonging degree.

Algorithm :

1) *Let S = {S1, S2, ..., Sn} a set of sources*

2) *Let G={G1,G2, ……, Gm}a set of groups generated by the distribution algorithm.*

3) *For each element of untreated group G_i, we proceed to the following iterations:*

a. Mark up the group G_i and calculate LowCoupling set (G_i).

b. If the set LowCoupling(G_i) is not empty then:

c. For each sources S_i in LowCoupling (G_i), do:

1) *Calculate the DA (Si, Gj), for all other groups Gj such that i # j, and store them in a indexed table by the groups $G_{i\,:}$ TabDegre[G_i] in descending order.*

2) *Traverse the table TabDegre from the first element until the verification of the condition:*

3) *TabDegre[G_k]> threshol (S_i,G_k)*

4) *If any group G_k from the table TabDegre don't validate this condition, we will create a new marked group Gn that contains the source S_i.*

5) *If not, we add Si in the group G_k.*

V. Study and Evaluation Of Network Overload

Generally, the aim of using the data warehouse is to ensure access to data in a distributed environment and minimizing network overhead. In this section, we study the performance of data transmission on the network during the interrogation of the mediator with the presence of local data warehouses generated by the algorithm that we proposed in the previous section. We will use analytical modeling and statistical analysis of simulation results. In particular, we examine the statistics of the packets transmission on the network, and we propose a comparative study on the distribution of network load among the proposed solutions. We then establish the relationship between the data warehouse system efficiency of data replication, which could be used to adjust dynamically the degree of replication depending on the bandwidth of the network, optimizing the tradeoff between storage and data accessibility. To do this, we consider the following parameters:

$\mathrm{Taille\_Reponse}(S_i)$: The average size of a response to a request asking the source S_i.

NbrSources : Total number of sources.

Taille_Rep_Moy(Ei): the average size of a response to a request asking the data warehouse Ei.

Nbr_Sources(E_i): The number of sources comprising data warehouse E_i.

Nbr_Entrepot : The number of data warehouses.

Nbr_Requete : Number of queries .

P(S_i): The probability to have a new response from the source S_i.

Let R a user query and $\{R_1, R_2, \dots, R_n\}$ a set of subqueries after rewriting by mediator and n sources {S_1, ……., S_n}.

A. *Without using classification methods*

In this case we assume that the sources are integrated by the mediator are independent, and for each source S_i, the mediator generates a subquery R_i.

$$\mathrm{TailleReponse}(R_1) = TailleRepMoy(S_1) \times \mathrm{P}(\mathrm{S}_1)$$

$$\mathrm{TailleReponse}(R_2) = \mathrm{TailleRepMoy}(\mathrm{S}_1) \times (\mathrm{P}(\mathrm{S}_2) - \mathrm{P}(\mathrm{S}_1))$$

$$\mathrm{TailleReponse}(R_3) = TailleRepMoy(S_3) \times \left(P(S_3) - P(S_1 \cap S_2\,)\right)$$

$$T\mathrm{ailleReponse}(\mathrm{R_n}) = \mathrm{TailleRepMoy}(\mathrm{S_n}) \times \left(\mathrm{P}(\mathrm{S_n}) - \mathrm{P}\left(\bigcap_{\mathrm{i=1}}^{\mathrm{n-1}} \mathrm{S_i}\right)\right)$$

$$T a\mathrm{illeReponse}(\mathrm{R_n}) = \mathrm{TailleRepMoy}(\mathrm{S_n}) \times \left(\mathrm{P}(\mathrm{S_n}) - \prod_{\mathrm{i=1}}^{\mathrm{n-1}} \mathrm{P}(\mathrm{S_i})\right)$$

$$T\mathrm{aille}(\mathrm{R}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{TailleReponse}(\mathrm{R_i})$$

$$T\mathrm{aille}(\mathrm{R}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{TailleRepMoy}(\mathrm{S_i}) \times \left(\mathrm{P}(\mathrm{S_i}) - \prod_{\mathrm{k=1}}^{\mathrm{i-1}} \mathrm{P}(\mathrm{S_k})\right)$$

For reasons of simplicity, we assume that the probability $\mathrm{P}(\mathrm{S_i})$ and average response size TailleRepMoy ($\mathrm{S_i}$) identical for all sources$\mathrm{S_i}$.

We represent this parameters respectively by P, and TM then :

$$\mathrm{Taille}(R) = \sum_{i=1}^{n} TM \times \left(P - \prod_{k=1}^{i-1} P\right)$$

$$\mathrm{Taille}(R) = TM \times \sum_{i=1}^{n-1} \left(P - P^{i-1}\right)$$

B. *With using classification methods*

In this case we consider another data duplication factor $\mathrm{D}(\mathrm{E_i})$ in a data warehouse $\mathrm{E_i}$. This factor represents the probability of data duplication in the responses of sources. Therefore:

$$P(E_i) = D(E_i) \times \prod_{j=1}^{Nbr(E_j)} P(S_j)$$

With, $\mathrm{Nbr}(\mathrm{E_i})$ is the number of sources of data warehouse $\mathrm{E_i}$. Also, we suppose K data warehouse generated after applying one of the classification algorithms. The overall size of the result after executing a query R is:

$$TailleReponse(R_1) = TailleRepMoy(E_1) \times P(E_1)$$

$$TailleReponse(R_2) = TailleRepMoy(E_1) \times (P(E_2) - P(E_1))$$

$$TailleReponse(R_3) = TailleRepMoy(E_3) \times \left(P(E_3) - P(E_1 \cap E_2\,)\right)$$

$$TailleReponse(R_k) = TailleRepMoy(E_k) \times \left(P(S_k) - P\left(\bigcap_{i=1}^{k-1} S_i\right)\right)$$

$$TailleReponse(R_k) = TailleRepMoy(E_k) \times \left(P(S_k) - \prod_{i=1}^{k-1} P(E_i)\right)$$

$$Taille(R) = \sum_{p=1}^{k} TailleReponse\left(R_p\right)$$

$$= \sum_{p=1}^{k} \mathrm{TailleRepMoy}\left(E_p\right) \times \left(P\left(E_p\right) - \prod_{i=1}^{p-1} P(E_i)\right)$$

$$= \sum_{p=1}^{k} \text{TailleRepMoy}(E_p) \times \left(\left(D(E_p) \times \prod_{i=1}^{Nbr(E_p)} P(S_i) \right) - \prod_{i=1}^{p-1} \left(D(E_i) \times \prod_{j=1}^{Nbr(E_i)} P(S_j) \right) \right)$$

(F1)

For the sake of simplicity, we assume that the probability $P(S_i)$, the replication factor $D(E_i)$ and the size of the average response TailleRepMoy (S_i) are regular for all sources. We represent these parameters respectively by P, D, and T then:

$$Taille(R) = \sum_{p=1}^{k} TMRE \times \left(\left(D \times \prod_{i=1}^{Nbr(E_p)} P \right) - \prod_{i=1}^{P-1} \left(D \times \prod_{j=1}^{Nbr(E_i)} P \right) \right)$$

$$Taille(R) = TMRE \times \sum_{p=1}^{k} \left(\left(D \times P^{Nbr(E_p)} \right) - \left(D^{p-1} \times P^{Nbr(E_i) \times (p-1)} \right) \right)$$

(F2)

1) Analysis:

According to the two formulas (F1) and (F2) the size of response to a query, it can be concluded that the size of data exchange on the network with the use of classification methods, following a series of interrogation of the mediator, is lower than without the use of classification methods in different situations. But the degree of difference depends on the factor of duplication, the average size of the query result, the number of remote sources and number of data warehouses generated by classifiers.

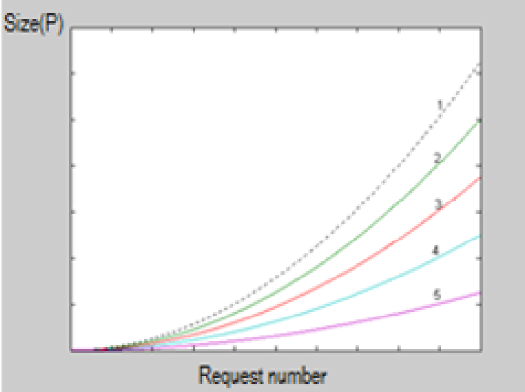


Fig .2. Influence of the probability P

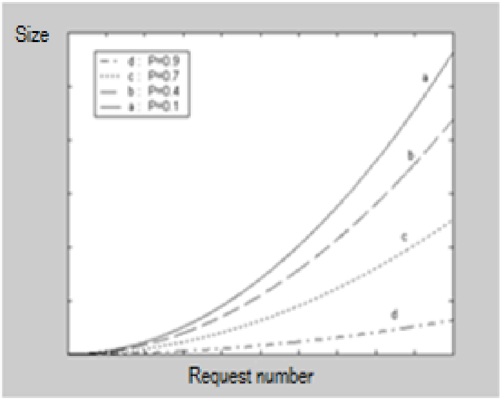


Fig .3. Influence of the duplication factor D

Generally, in the case of use of classification methods it can be seen that the rate of communication and exchange of data decreases to a certain level of interrogation. Therefore, the cost estimate takes into account other additives parameters influence on the basic parameters studied previously.

For example, we assumed that the probabilities, P and D are constant for the any new responses from a remote source regardless of the number K of warehouse generated by the classification algorithms. But these probabilities depend heavily on this number K and the number of queries N.

The average size of the response decreases for each new query. This degradation depends mainly to the identical records of sources in a warehouseE_i.

Indeed, in the first experiment, we fixe two parameters: the duplication factor D and the average number of warehouses K, and we changed the number of queries N. The results are shown in the following figure.

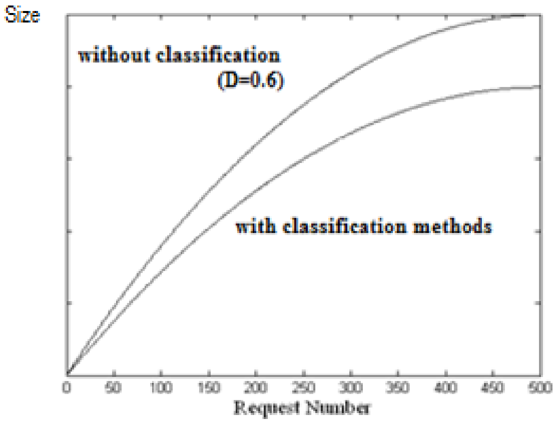


Fig .4. Comparison between methods (D=0.6)

According to this graph, the size of the exchanges on the network without the use of classification methods is always higher than that with the use of classification methods. The difference becomes important with the increase of the number of questions. This means that duplication of data stored in the warehouse (as the duplication factor D) influence on the exchange rate.

In the second experiment we can observe the effect of the variation of the duplication factor D on the exchange rate. According to figure 5, we note that if the duplication factor D decreases, the number of warehouses increases, therefore the exchange rate also increases. For D = 0, this meant that there is no classification groups. This shows that the classification methods guarantee a better system performance.

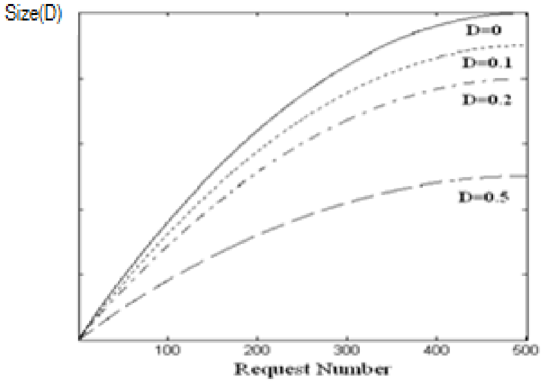


Fig .5. Influence of the duplication factor D on the size.

VI. Conclusion

In this paper, we have presented a new approach for query optimization using dynamic programming technique for set integrated heterogeneous sources. To do this, we have developed a relevant algorithm which grouping the sources into subsets based on our new classification method that we proposed in this paper. In fact, we have shown with the study of the performance of data transmission on the network during the interrogation of the mediator with the presence of local data warehouses generated by our proposed algorithm and the evaluation of network overload that our classification methods using datawarehousing offer many benefits: minimized the cost of data exchange during the execution of subqueries on geographically remote sites, increased flexibility, and simplified access to data with greater agility. Indeed, local data warehousing offers space and interrogation power for several sources centers, but also the possibility of analyzing the data more efficiently on any remote server based on availability and needs. This solution effectively enables more users to access to more and more data without difficulty. Thanks to the Distributed Databases solution, we can migrate critical data on a data centers and improve the response time of readjustment and equilibration of the distribution data localization. In the perspective, we will study the performance of the different solutions by a comparative study.

Refrences

[1] K. Asghari, A S. Mamaghani and M R. Meybodi, “An Evolutionary Approach for Query Optimization Problem in Database”, In Proc. of Int. Joint Conf. on Computers, Information and System Sciences, and Engineering (CISSE2007), England, springer, 2007.

[2] L. Cherrat, M. Ezziyyani and M. Essaaidi, “Automatic Generation of Query Order Execution Plan for Hybrid Mediator with Medical Sources”, In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), International Conference on Digital Object Identifier, Beijing, pp. 558 - 565, 10-12 Oct. 2011.

[3] A. Cali, D. Calvanese, G. Giacomo and M. Lenzerini, “ On the Expressive Power of Data Integration Systems”, In Proc. of the 21st Inter. Conf. on Conceptual Modeling, pp.338-350, 2003.

[4] M. Ezziyyani, M. Bennouna and L.Cherrat, “Mediator of the heterogeneous information systems based on application domains specification : AXMed Advanced XML Mediator”, IEEE Journal,, Vol.3, No.2, pp. 25-45, 2006.

[5] L M. Haas, “Beauty and the beast: The theory and practice of information integration”, In Database Theory ICDT'2007, pp.28-43, 2007.

[6] D. Calvanese, and G D. Giacomo, “Data Integration: A Logic-Based Perspective”, AI Magazine, Vol.26, No.1, 2005.

[7] I. Jellouli, M. El Mohajir and E. Zimanyi, “Classification conceptuelle et ontologie de domaine pour l’intégration sémantique des données”, La revue électronique des technologies d'information e-TI, No. 5, 5 novembre 2008.

[8] S. Kermanshahani, “Semi-materialized framework: a hybrid approach to data integration”, In Proc. of the 5th inter. conf. on Soft computing as transdisciplinary science and technology CSTST '08, pp. 600-606, New York, NY, USA, 2008.

[9] A. Halevy, A. Rajaraman. and J. Ordille, “Data integration: the teenage years”, Proceedings of the 32nd international conference on Very large data bases VLDB '06, pp. 9-16, Seoul, Korea, September 12-15, 2006.

[10] M S. Hacid, and C. Reynaud, “L'intégration de sources de données”, Revue Information - Interaction - Intelligence I3, Vol.4, No. 2, 2004.

[11] Z G. Ives, AY. Levy, D S. Weld, D. Florescu and M.. Friedman, “Adaptive Query Processing for Internet Applications”, In IEEE Computer Society Journal, Vol.23, No. 2, pp. 19-26, June 2000.

[12] Z G. Ives, “Efficient query processing for data integration”, Doctoral thesis at the University of Washington, 2002.

[13] H P. Kriegel, P. Kunath, M. Pfeifle and M. Renz, “Approximated Clustering of Distributed High-Dimensional Data”, In Proc. of the 9 th Pacific-Asia conference (PAKDD 2005), Hanoi, Vietnam, pp. 432-441, May 18-20, 2005.

[14] E Johnson. and H. Kargupta, “Hierarchical Clustering From Distributed, Heterogeneous Data”, In Computer Science Journal, pp. 221-244. Springer-Verlag, 1999.

[15] A K. Jain., M N. Murty and P J. Flynn, “Data Clustering: A Review. In ACM Computing Surveys”, Vol. 31, No. 3, pp. 265-323, Sep. 1999.

[16] N F. Samatova, G. Ostrouchovand, A. Geist and A V. Melechko, “RACHET: An Efficient Cover-Based Merging of Clustering

Hierarchies from Distributed Datasets", In Distributed and Parallel Databases, Vol. 11, No.2, pp. 157-180, Mars 2002.

[17] E. Januzaj, H P. Kriegel, and M. Pfeifle, "DBDC: Density Based Distributed Clustering", In Proc. 9th Int. Conf. on Extending Database Technology (EDBT 2004), pp. 88-105, Heraklion, Greece, 2004.

[18] E. Januzaj, H P. Kriegel and M. Pfeifle, "Scalable Density-Based Distributed Clustering", In Proc. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa, Italy, 2004.

[19] H P. Kriegel, S. Brecheisen, P. Kröger, M. Pfeifle and M. Schubert, "Using Sets of Feature Vectors for Similarity Search on Voxelized CAD Objects", In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'03), pp.587-598, San Diego, CA, 2003.

[20] M C. Tu, D. Shin and D.Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", In Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing , pp.183-187, 2009.

[21] A. S. Kumar and S. Sahni, "Comparative Study of Classification Algorithms for Spam Email Data Analysis", In International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 5, May 2011.

[22] P. Berkhin, "A Survey of Clustering Data Mining Techniques In Grouping Multidimensional Data", pp. 25-71, 2006.

[23] J P. Nakache and J. Confais , "Approche pragmatique de la classification", In livre editions TECHNIP, 2004.

[24] A. Lelu, "Evaluation de trois mésures de Similarité utilisées en Sciences de l'information"., In Information Sciences for Decision Making , Vol.6, pp.14-25, 2003.

[25] S H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions", International Journal Of Mathematical Models And Methods In Applied Sciences, Vol. 1, No. 42007.

[26] S H. Cha, "Taxonomy of Nominal Type Histogram Distance Measures", In American Conference On Applied Mathematics (Math '08), Harvard, Massachusetts, USA, March 24-26, 2008.

A Hybrid Framework using RBF and SVM for Direct Marketing

M.Govidarajan
Assistant Professor
Department of Computer Science and Engineering
Annamalai University
Annamalai Nagar-608002
Tamil Nadu, India

***Abstract*—one of the major developments in machine learning in the past decade is the ensemble method, which finds highly accurate classifier by combining many moderately accurate component classifiers. This paper addresses using an ensemble of classification methods for direct marketing. Direct marketing has become an important application field for data mining. In direct marketing, companies or organizations try to establish and maintain a direct relationship with their customers in order to target them individually for specific product offers or for fund raising. A variety of techniques have been employed for analysis ranging from traditional statistical methods to data mining approaches. In this research work, new hybrid classification method is proposed by combining classifiers in a heterogeneous environment using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. Here, modified training sets are formed by resampling from original training set; classifiers constructed using these training sets and then combined by voting. Empirical results illustrate that the proposed hybrid systems provide more accurate direct marketing system.**

Keywords—Direct Marketing; Ensemble; Radial Basis Function; Support Vector Machine; Classification Accuracy.

I. Introduction

Data mining methods may be distinguished by either supervised or unsupervised learning methods. In supervised methods, there is a particular pre-specified target variable, and they require a training data set, which is a set of past examples in which the values of the target variable are provided.

Direct marketing [20] has become an important application field for data mining. In direct marketing [2] companies or organizations try to establish and maintain a direct relationship with their customers in order to target them individually for specific product offers or for fund raising. Large databases of customer and market data are maintained for this purpose. The customers or clients to be targeted in a specific campaign are selected from the database, given different types of information such as demographic information and information on the customer's personal characteristics like profession, age and purchase history.

The customers of a company are regarded as valuable business resources in competitive markets, leading to efforts to systematically prolong and exploit existing customer relations. Consequently, the strategies and techniques of customer relationship management (CRM) have received increasing attention in management science. CRM features data mining as a technique to gain knowledge about customer behaviour and preferences.

Data mining problems in the CRM domain, such as response optimization to distinguish between customers who will react to a mailing campaign or not, chum prediction, in the form of classifying customers for churn probability, cross-selling, or up-selling are routinely modeled as classification tasks, predicting a discrete, of- ten binary feature using empirical, customer centered data of past sales, amount of purchases, demographic or psychographic information etc. Customer retention has a significant impact on firm profitability. Gupta et al find that a 1% improvement in retention can increase firm value by 5%. [9]. Churn refers to the tendency for customers to defect or cease business with a company. Marketers interested in maximizing lifetime value realize that customer retention is a key to increasing long-run firm profitability. A focus on customer retention implies that firms need to understand the determinants of customer defection (churn) and are able to predict those customers who are at risk of defection at a particular point in time.

Response modeling is usually formulated as a binary classification problem. The customers are divided into two classes, respondents and non-respondents. A classifier is constructed to predict whether a given customer will respond or not. From a modeling point of view, however, several difficulties arise [22] [28]. One of the most noticeable is a severe class imbalance resulting from a low response rate: typically less than 5% of customers are respondents [7]. A typical binary classifier will result in lopsided outputs to the non-respondent class [14].

In other words, the classifier will predict most or even all customers not to respond. Although the classification accuracy may be very high since a majority of customers are in fact non-respondents. In this work, a model which identifies a subset of customers is constructed that includes as many respondents and as few non-respondents as possible Various classification methods have been used for response modeling such as statistical and machine learning methods. Recently, SVMs have drawn much attention and a few researchers have implemented them for response modeling [22] [27].

Classification is a very common data mining task. In the process of handling classification tasks, an important issue usually encountered is determining the best performing method for a specific problem [13]. Several studies address the issue. For example, Michie, Spiegelhalter, and Taylor try to find the relationship between the best performing method and data types of input/output variables.[18] Hybrid models have been suggested to overcome the defects of using a single supervised learning method, such as radial basis function and support vector machine techniques. Hybrid models combine different methods to improve classification accuracy. The term combined model is usually used to refer to a concept similar to a hybrid model. Combined models apply the same algorithm repeatedly through partitioning and weighting of a training data set. Combined models also have been called Ensembles. Ensemble improves classification performance by the combined use of two effects: reduction of errors due to bias and variance [11]. Recently, hybrid data mining approaches have gained much popularity; however, a few studies have been proposed to examine the performance of hybrid data mining techniques for response modeling [17].

This paper proposes a new hybrid classification method to improve the Classification accuracy. The primary objective of this paper is to construct ensemble of radial basis function and Support Vector Machine is to predict whether a given customer will respond or not for direct marketing in terms of classification accuracy.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents hybrid direct marketing system and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

II. Related Work

Direct marketing aims at obtaining and maintaining direct relations between suppliers and buyers within one or more product/market combinations. In marketing, there are two main different approaches to communication: mass marketing and direct marketing [16]. Mass marketing uses mass media such as print, radio and television to the public without discrimination. While direct marketing involves the identification of customers having potential market value by studying the customers' characteristics and the needs (the past or the future) and selects certain customers to promote. Direct marketing becomes increasingly popular because of the increased competition and the cost problem. It is an important area of applications for data mining, data warehousing, statistical pattern recognition, and artificial intelligence. In direct marketing, models (profiles) are generated to select potential customers (from the client database) for a given product by analyzing data from similar campaigns, or by organizing test mail campaigns [21]. Various classifiers have been employed such as logistic regression, neural networks and support vector machine.

Aristides Gionis et al have shown that the numbers of clusters discovered by their algorithms seem to be very reasonable choices: for the Votes dataset most people vote according to the official position of their political parties, so having two clusters is natural; for the Mushrooms dataset, notice that both ROCK and LIMBO achieve much better. [1]. Many aspects of churn have been modeled in the literature. First, whether churn is hidden or observable influence the overall approach to modeling. In some industries, customer defection is not directly observed, as customers do not explicitly terminate a relationship, but can become inactive. In other industries, however, the defection decision is observable as customers cease their relationship via actively terminating their contract with the firm [4].

The modeling approach could also depend critically on the relative importance placed on explanation/interpretation *vis a vis* prediction. Models that are better at explanation may not necessarily be better at prediction. The empirical literature in marketing has traditionally favored parametric models (such as logistic or probit regression or parametric hazard specifications and zero-inflated poisson models) that are easy to interpret. Similar to the previous discussion on acquisition, churn is a rare event that may require new approaches from data mining, machine learning and non-parametric statistics that emphasize predictive ability [10]. These include projection-pursuit models, jump diffusion models, neural network models, tree structured models, spline-based models such as Generalized Additive Models (GAM), and Multivariate Adaptive Regression Splines (MARS), and more recently approaches such as support vector machines and boosting [15].

Tang applied feed forward neural network to maximize performance at desired mailing depth in direct marketing in cellular phone industry. He showed that neural networks show more balance outcome than statistical models such as logistic regression and least squares regression, in terms of potential revenue and churn likelihood of a customer [23].

Xu et al proposed four combining classifier approaches according to the levels of information available from the various classifiers. The experimental results showed that the performance of individual classifiers could be improved significantly. [26]

Freund and Schapire proposed an algorithm the basis of which is to **a**daptively **r**esample and **c**ombine (hence the acronym--arcing) so that the weights in the resampling are increased for those cases most often misclassified and the combining is done by weighted voting. [5] [6]. Previous work has demonstrated that arcing classifiers is very effective for RBF-SVM hybrid system. [8]. It is surprising there are only few papers that seek to assess the state of research in this area, or outline the challenges unique to this area. This paper seeks to address this void.

In this paper, a direct marketing system is proposed using radial basis function and support vector machine and the effectiveness of the proposed RBF-SVM hybrid system is evaluated by conducting several experiments on voting database. The performance of the RBF-SVM hybrid classifier is examined in comparison with standalone RBF and standalone SVM classifier. This work focuses to understand the relative merits of the base and proposed hybrid approaches for CRM applications.

III. Hybrid Direct Marketing System

This section shows the proposed RBF-SVM hybrid system which involves Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers.

A. *RBF-SVM Hybrid System*

The proposed hybrid direct marketing system is composed of three main phases; preprocessing phase, classification phase and combining Phase.

1) Voting Dataset Preprocessing: First the data is collected from the United States Congressional Voting Records Database. Before performing any classification method the data has to be preprocessed. In the data preprocessing stage it has been observed that the datasets consist of many missing value attributes. By eliminating the missing attribute records may lead to misclassification because the dropped records may contain some useful pattern for Classification. The dataset is preprocessed by removing missing values using supervised filters.

2) Existing Classification Methods

*a) Radial Basis Function Neural Network:*The RBF [19] design involves deciding on their centers and the sharpness (standard deviation) of their Gaussians. Generally, the centres and SD (standard deviations) are decided first by examining the vectors in the training data. RBF networks are trained in a similar way as MLP. The output layer weights are trained using the delta rule. The RBF networks used here may be defined as follows.

- RBF networks have three layers of nodes: input layer, hidden layer, and output layer.
- Feed-forward connections exist between input and hidden layers, between input and output layers (shortcut connections), and between hidden and output layers. Additionally, there are connections between a bias node and each output node. A scalar weight is associated with the connection between nodes.
- The activation of each input node (fanout) is equal to its external input where is the th element of the external input vector (pattern) of the network (denotes the number of the pattern).
- Each hidden node (neuron) determines the Euclidean distance between “its own” weight vector and the activations of the input nodes, i.e., the external input vector the distance is used as an input of a radial basis function in order to determine the activation of node. Here, Gaussian functions are employed. The parameter of node is the radius of the basis function; the vector is its center.
- Each output node (neuron) computes its activation as a weighted sum The external output vector of the network, consists of the activations of output nodes, i.e., The activation of a hidden node is high if the current input vector of the network is “similar” (depending on the value of the radius) to the center of its basis function. The center of a basis function can, therefore, be regarded as a prototype of a hyper spherical cluster in the input space of the network. The radius of the cluster is given by the value of the radius parameter.

b) Support Vector Machine:

The support vector machine (SVM) is a recently developed technique for multi dimensional function approximation. The objective of support vector machines is to determine a classifier or regression function which minimizes the empirical risk (that is the training set error) and the confidence interval (which corresponds to the generalization or test set error) [24].

Given a set of N linearly separable training examples $\mathbf{S} = \{\mathbf{x_i} \in \mathbf{R^n} | \mathbf{i} = \mathbf{1,2,...,N}\}$, where each example belongs to one of the two classes, represented by $y_i \in \{\div 1, -1\}$, the SVM learning method seeks the optimal hyperplane **w . x** +b = 0, as the decision surface, which separates the positive and negative examples with the largest margins. The decision function for classifying linearly separable data is:

$$f(\mathbf{X}) = sign(W.X + b) \qquad (1)$$

Where **w** and b are found from the training set by solving a constrained quadratic optimization problem. The final decision function is

$$\mathrm{f(x) = sign}\left(\sum_{i=1}^{N} \mathrm{a_i y_i (x_i..x) + b}\right) \qquad (2)$$

The function depends on the training examples for which a_is is non-zero. These examples are called support vectors. Often the number of support vectors is only a small fraction of the original data set. The basic SVM formulation can be extended to the non linear case by using the nonlinear kernels that maps the input space to a high dimensional feature space. In this high dimensional feature space, linear classification can be performed. The SVM classifier has become very popular due to its high performances in practical applications such as text classification and pattern recognition. The support vector regression differs from SVM used in classification problem by introducing an alternative loss function that is modified to include a distance measure. Moreover, the parameters that control the regression quality are the cost of error C, the width of tube $\boldsymbol{\varepsilon}$ and the mapping function $\boldsymbol{\phi}$.

In this research work, the values for polynomial degree will be in the range of 0 to 5. In this work, best kernel to make the prediction is polynomial kernel with epsilon = 1.0E-12, parameter d=4 and parameter c=1.0. A hybrid scheme based on coupling two base classifiers using arcing classifier adapted to data mining problem is defined in order to get better results.

3) Proposed RBF-SVM Hybrid System

Given a set D, of d tuples, arcing works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset *D* will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples d from training dataset D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The hybrid classifier (RBF-SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid RBF-SVM using Arcing Classifier
Input:

- D, a set of d tuples.
- k = 2, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine)

Output: Hybrid RBF-SVM model, M^*.
Procedure:

```
1. For i = 1 to k do // Create k models
2.    Create a new training dataset, $D_i$, by sampling D with replacement. Same example from given dataset $D$ may occur more than once in the training dataset $D_i$.
3.    Use $D_i$ to derive a model, $M_i$
4.    Classify each example d in training data $D_i$ and initialized the weight, $W_i$ for the model, $M_i$, based on the accuracies of percentage of correctly classified example in training data $D_i$.
5. endfor
```

To use the hybrid model on a tuple, X:

```
1. if classification then
2.    let each of the k models classify X and return the majority vote;
3. if prediction then
4.    let each of the k models predict a value for X and return the average predicted value;
```

The basic idea in Arcing [3] is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

IV. Performance Evaluation Measures

A. Cross Validation technique

Cross-validation (Jiawei Han and Micheline Kamber**,** 2003) sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds.

B. Criteria for Evaluation

The primary metric for evaluating classifier performance is classification Accuracy: the percentage of test samples that the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

V. Experimental Results and Discussion

A. Dataset Description

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

B. Experiments and Analysis

An experimental evaluation of the competing methods is conducted in the domain of CRM, striving to exemplify the adequacy and performance of RBF versus SVM versus proposed hybrid RBF-SVM for the task of response optimization in terms of accuracy based upon an numerical experiment. The voting dataset are taken to evaluate the proposed RBF-SVM direct marketing system. All experiments have been performed using Intel Core 2 Duo 2.26 GHz processor with 2 GB of RAM and weka software [25].

TABLE I. THE PERFORMANCE OF BASE AND HYBRID CLASSIFIERS

Dataset	Classifiers	Classification Accuracy
Voting dataset	RBF	94.48 %
	SVM	96.09 %
	Proposed Hybrid RBF-SVM	99.31 %

The data set described in section 5 is being used to test the performance of base classifiers and hybrid classifier. Classification accuracy was evaluated using 10-fold cross validation. In the proposed approach, first the base classifiers RBF and SVM are constructed individually to obtain a very good generalization performance. Secondly, the ensemble of RBF and SVM is designed. In the ensemble approach, the final output is decided as follows: base classifier's output is given a weight (0–1 scale) depending on the generalization performance as given in Table 1.

The results of the computational experiments are presented in Table 1, comparing the performance of RBF, SVM and proposed hybrid RBF-SVM models on the generalization set. For the case of response optimization the accuracy is of primarily importance, as it measures the amount of correctly classified respondents. The accuracy of SVM was found to be higher than RBF classifier and the proposed hybrid RBF-SVM exhibits higher percentage accuracy than the individual classifiers. Thus the proposed hybrid RBF-SVM model can be regarded as very good for the application domain.

According to Table 1, the proposed hybrid model shows significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant. The x^2 statistic x2 is determined for all the above approaches and their critical value is found to be less than 0.455. Hence corresponding probability is $p < 0.5$. This is smaller than the conventionally accepted significance level of 0.05 or 5%. Thus examining a x^2 significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of x^2 statistic analysis shows that the proposed classifiers are significant at $p < 0.05$ than the existing classifiers.

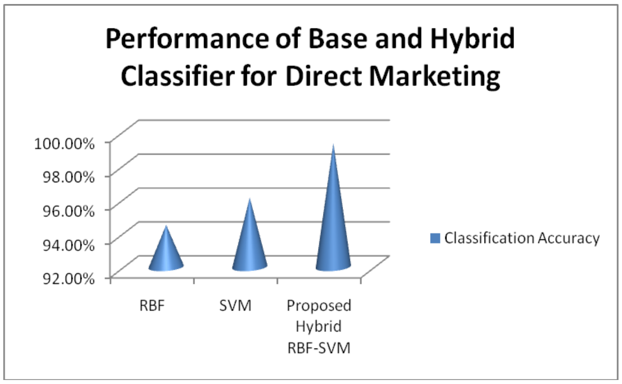


Fig. 1. Classification Accuracy

The experimental results show that proposed ensemble of RBF and SVM is superior to individual approaches for direct marketing problem in terms of Classification accuracy.

VI. Conclusion

In this research, some new techniques have been investigated for direct marketing and their performance is evaluated based on the Voting dataset. Recently, various architectures from computational intelligence and machine learning, such as artificial neural networks (ANN) and support vector machines (SVM) have found increasing consideration in practice, promising effective and efficient solutions for classification problems in real-world applications through robust generalization in linear and non-linear classification problems, deriving relationships directly from the presented sample data without prior modeling assumptions. Hence RBF and SVM are explored as direct marketing models. Next a hybrid RBF-SVM model is designed using RBF and SVM models as base classifiers. Thus, a hybrid intelligent direct marketing system is proposed to make optimum use of the best performances delivered by the individual base classifiers and the hybrid approach. The hybrid RBF-SVM shows higher percentage of classification accuracy than the base classifiers. This paper provides some insights on the relative performance of base and hybrid approaches to predictive modeling for churn based on classification accuracy for modeling defections.

The numerical results show, that RBF and SVM are both suitable for the task of response optimization, leading to classification accuracy that can be considered as very good for practical problems. This robustness makes SVM best suited for users who are less experienced in data mining and model building, which is not untypical in business environments. Consequently, the hybrid RBF-SVM is recommended in standard data mining software packages like WEKA as the proposed technique is easy to manage and provides competitive results. Finally, early detection and prevention of customer attrition can also enhance the total lifetime of the customer base, if efforts are focused on the retention of valuable customers. The future research will focus on investigate other ways to combine basic models in order to create more accurate models in response modeling and therefore, minimize marketing expenses and bring in more profits to the company.

Acknowledgment

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work.

References

[1] Aristides Gionis and Heikki Mannila and Panayiotis Tsaparas. (2005), Clustering Aggregation. ICDE.

[2] C. L. Bauer (1998). A direct mail customer purchase model, Journal of Direct Marketing, 2:16–24.

[3] Breiman. L, (1996), “Bias, Variance, and Arcing Classifiers”, Technical Report 460, Department of Statistics, University of California, Berkeley, CA.

[4] Fader, P. S., B. G. S. Hardie, and K. L. Lee. (2004). “Counting Your Customers’ the Easy Way: An Alternative to the Pareto/NBD Model,” Working Paper, Wharton Marketing Department.

[5] Freund, Y. and Schapire, R. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In proceedings of the Second European Conference on Computational Learning Theory, pp 23-37.

[6] Freund, Y. and Schapire, R. (1996) Experiments with a new boosting algorithm, In Proceedings of the Thirteenth International Conference on Machine Learning, 148-156 Bari, Italy.

[7] Gonul, F. F., Kim, B. D., & Shi, M. (2000). Mailing smarter to catalog customers. Journal of Interactive Marketing, 14(2), 2–16.

[8] M.Govindarajan, RM.Chandrasekaran, (2012), “Intrusion Detection using an Ensemble of Classification Methods”, In Proceedings of International Conference on Machine Learning and Data Analysis, pages 459-464.

[9] Gupta, Sunil, Donald R. Lehmann, and Jennifer Ames Stuart. (2004). “Valuing Customers,” Journal of Marketing Research 41(1), 7–18.

[10] Hastie, T., R. Tibshirani, and J. Friedman. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag.

[11] Haykin, S. (1999). Neural networks: a comprehensive foundation (second ed.). New Jersey: Prentice Hall.

[12] Jiawei Han , Micheline Kamber**,** (2003), “ Data Mining – Concepts and Techniques” Elsevier Publications.

[13] Joon Hur, Jong Woo Kim, (2008), “A hybrid classification method using error pattern modeling”, Expert Systems with Applications, 34, 231–241.

[14] Kubat, M., Holte, R., & Matwin, S. (1997). Learning when negative examples abound. Lecture Notes in Artificial Intelligence (LNAI 1224), 146–153, Prague, The Czech Republic

[15] Lemmens, Aur´elie and Christophe Croux. (2003). “Bagging and Boosting Classification Trees to Predict Churn”, Working Paper, Teradata center.

[16] Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions Proceedings of the KDD98 pp. 73–79.

[17] Maryam Daneshmandi, Marzieh Ahmadzadeh (2013), “A Hybrid Data Mining Model to Improve Customer Response Modeling in Direct Marketing, Indian Journal of Computer Science and Engineering, Vol. 3 No.6, 844-855.

[18] Michie, D., Spiegelhalter, D. J., & Taylor, C. (1994). Machine learning. Neural and statistical classification. Ellis Horwood.

[19] Oliver Buchtala, Manuel Klimek, and Bernhard Sick, Member, IEEE, (2005) **"**Evolutionary Optimization of Radial Basis Function Classifiers for Data Mining Applications", IEEE Transactions on systems, man, and cybernetics—part b: cybernetics, vol. 35, no. 5.

[20] Sara Madeira Joao M.Sousa (2000), "Comparison of target selection methods in direct Marketing" Technical University of Lisbon, Institution Superior Technician, Dept. Mechanical Eng./IDMEC, 1049-001 Lisbon, Portugal.

[21] Setnes, M., & Kaymak, U. (2001). Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing. IEEE Transactions on Fuzzy Systems, 9(1), 153–163.

[22] Shin, H. J., & Cho, S. (2006). Response modeling with support vector machines. Expert Systems with Applications, 30(4), 746–760.

[23] Tang, Z. (2011). "Improving Direct Marketing Profitability with Neural Networks." International Journal of Computer Applications 29(5): 13-18.

[24] Vapnik, V. (1998). Statistical learning theory, New York, John Wiley & Sons.

[25] Weka: Data Mining Software in java http://www.cs.waikato.ac.nz/ml/weka/

[26] L. Xu, A. Krzyzak, and C. Y. Suen, (1992), "Methods of Combining Multiple Classifiers and Their Applications to Handwritten Recognition", IEEE Transactions on Systems, Man, Cybernetics, Vol. 22, No. 3, pp. 418-435.

[27] Yu, E., & Cho, S. (2006). Constructing response model using ensemble based on feature subset selection. Expert Systems with Applications, 30(2), 352–360.

[28] Zahavi, J., & Levin, N. (1997). Issues and problems in applying neural computing to target marketing. Journal of Direct Marketing, 11(4), 63–75.

AUTHOR PROFILE

M.Govindarajan received the B.E and M.E and Ph.D Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001 and 2005 and 2010 respectively. He did his post-doctoral research in the Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom in 2011. He is currently an Assistant Professor at the Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He has presented and published more than 70 papers in Conferences and Journals. His current research interests include Data Mining and its applications, Web Mining, Text Mining, and Sentiment Mining. He was the recipient of the Achievement Award for the field and to the Conference Bio-Engineering, Computer science, Knowledge Mining (2006), Prague, Czech Republic and All India Council for Technical Education "Career Award for Young Teachers (2006), New Delhi, India. He is active Member of various professional bodies.

Time Variant Change Analysis in Satellite Images

Rachita Sharma
Computer Science & Engg. Deptt.
GNIT Girls Institute of Technology
Greater Noida, India

Sanjay Kumar Dubey
Assistant Professor, Computer Science & Engg. Deptt.
Amity University, Uttar Pradesh
Sec. 125, Noida, India

Abstract—**This paper describes the time variant changes in satellite images using Self Organizing Feature Map (SOFM) technique associated with Artificial Neural Network. In this paper, we take a satellite image and find the time variant changes using above technique with the help of MATLAB. This paper reviews remotely sensed data analysis with neural networks. First, we present an overview of the main concepts underlying Artificial Neural Networks (ANNs), including the main architectures and learning algorithms. Then, the main tasks that involve ANNs in remote sensing are described. We first make a brief introduction to models of networks, for then describing in general terms Artificial Neural Networks (ANNs). As an application, we explain the back propagation algorithm, since it is widely used and many other algorithms are derived from it. There are two techniques that are used for classification in pattern recognition such as Supervised Classification and Unsupervised Classification. In supervised learning technique the network knows about the target and it has to change accordingly to get the desired output corresponding to the presented input sample data. Most of the previous work has already been done on supervised classification. In this study we are going to present the classification of satellite images using unsupervised classification method of ANN.**

Keywords*—*Satellite Images; SOFM; ANN; Supervised Classification, Unsupervised classification.

I. INTRODUCTION

Time series is a chain of observations that take part according to the time like weekly series of newspaper, hourly observations of news happening in the city, number of criminal records etc. Most popular fields of time variant changes are business economics, remote sensing and weather forecasting [1]. Basically time variant changes are dependent on some current adjacent observations. In time variant changes the goal is to create a prototype that can reveal the current process and predict the future of the measured process by calculating the values of certain variables sequentially in time. Usually data is incomplete and also having noise. So we try to find exact data possible be compensated by adjustment of the input series value.

There are two statistical methods to gain temporal information of an image. First is linear i.e. AR and ARMA and second is nonlinear i.e. NARMAX and MARS [2]. In now a day neural network in time variant prediction is convert temporal sequence into concatenated vector via a tapped delay line and to feed the resulting vector as input to a network. Most recurrent neural networks are trained via supervised learning rules. However in temporal sequence analysis unsupervised network could reveal useful information from the temporal sequence at hand in analogy to unsupervised neural networks reported power in cluster analysis and dimensionality reduction [3]. In unsupervised learning or self organizing learning provision is made for task independent measure of the quality of representation that the network is required to learn, and free parameters of the network are optimized with respect to that measure, once the network has become tuned to form internal representation for encoding features of the input and thereby to create new classes automatically [4].

Artificial Neural Networks are the computing models that are inspired by biological neural network and provide new directions to solve problems arising in natural tasks. In particular, it is hoped that neural network would extract the relevant features from the input data and perform a pattern recognition task by learning from examples without explicitly stating the rules for performing the task.

Currently most of the neural networks models are severely limited in their abilities to solve real world problems. For problems such as speech recognition, image processing, natural language processing and decision-making, it is not normally possible to see a direct mapping of the given problem on to a neural network model. These are natural tasks, which human beings are good at, but we still do not understand how we do them. Hence it is a challenging task to find suitable neural network models to address these problems.

Automatic recognition, description, classification and grouping of patterns are important problems in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, artificial intelligence and remote sensing. In the most pattern recognition problems, patterns have a dynamic nature and non-adaptive algorithms (instruction sets) will fail to give a realistic solution to the problem. So in these cases, adaptive algorithms are used and among them, neural networks have the greatest hit. For example, the defenses applications very frequently need to record, detect, identify and classify images of objects or signals coming from various directions and from various sources- static or dynamic [5]. There are many applications in remote sensing like deforestation, effects of natural and manmade disasters, migration in the path of river due to dynamic nature of earth plates where study of dynamic data is needed.

Artificial Neural Networks (ANN) can play a role in such applications because of their capability to model nonlinear processes and to identify unknown patterns and images based on their learning model, or to forecast certain outcomes by extrapolation. On the basis of properties like steepness of slopes, local relief (the maximum local difference in elevation) and cross sectional forms of valley and divides, and texture of the surface material etc. considerable foresight can be achieved regarding temporal changes in land patterns. The land pattern of landform differences is strongly related in the arrangement of such other features of the natural environment as climate, soils, and vegetations.

In the present work we wish to classify satellite images using ANN's pattern recognition and classification capabilities. The Unsupervised Classification approach uses self organizing feature map to classify the patterns. The Self-organizing feature maps (SOFM) transform the input of arbitrary dimension into a one or two dimensional discrete map subject to a topological (neighborhood preserving) constraint.

The feature maps are computed using Kohonen unsupervised learning [6]. The output of the SOFM can be used as input to a supervised classification neural network such as the MLP. This network's key advantage is the clustering produced by the SOFM which reduces the input space into representative features using a self-organizing process. Hence the underlying structure of the input space is kept, while the dimensionality of the space is reduced.

II. Artificial Neural Network

The ANN is usually implemented using electronic components (digital or analog) and/or simulated on a digital computer. It employs massive interconnection of simple computing cells called 'neurons' or "processing elements (PE)". It resembles the brain in two ways: Knowledge is acquired by the network through learning process and Inter neuron connection strengths (synaptic weights) are responsible for storing the knowledge. The way the synaptic weights change is what makes the design of ANNs. An ANN works as follows*:*

A neuron receives inputs from a large number of other neurons or from an external stimulus. Weighted sums of these inputs are fed into a nonlinear activation function. The output of this function is fanned out (distributed) to connections to other neurons. The topology of neuron connections defines the flow of information in the network. The way the weights are adjusted in the network constitutes the learning process. Thus the three essential components of an ANN computational system are- activation function, architecture, and, the learning law.

Due to the differences in these three components, different ANN structures are explored for various applications and these structures differ in their computational complexities and requirements. The main attributes of neural processing are its nonlinear and adaptive learning capability, which enable machines to recognize possible variations of a same object or pattern and/or to identify unknown functions and mappings based on a finite set of training data, which can be noisy with missing information. Based on this 'Training by example' property with strong support of statistical and optimization theories, neural networks are becoming one of the most powerful and appealing nonlinear and adaptive data analysis tools for a variety of signal processing applications [7].

III. Classification OF Satellite Images Using SOFM

In this work self-organizing feature map network identifies a winning neuron using the same procedure as employed by a competitive layer. However, instead of updating only the winning neuron, all neurons within a certain neighborhood of the winning neuron are updated using the Kohonen rule.

The weights of the winning neuron (a row of the input weight matrix) are adjusted with the Kohonen learning rule. Supposing that the ith neuron wins, the elements of the ith row of the input weight matrix are adjusted as shown below [8].

$$iIW1,1(q) = iIW1,1(q-1)+\alpha(p(q)-iIW1,1(q-1))$$

The Kohonen rule [13] allows the weights of a neuron to learn an input vector, and because of this it is useful in recognition applications. Thus, the neuron whose weight vector was closest to the input vector is updated to be even closer. The result is that the winning neuron is more likely to win the competition the next time a similar vector is presented and less likely to win when a very different input vector is presented.

As more and more inputs are presented, each neuron in the layer closest to a group of input vectors soon adjusts its weight vector toward those input vectors. Eventually, if there are enough neurons, every cluster of similar input vectors will have a neuron that outputs 1 when a vector in the cluster is presented, while outputting a 0 at all other times. Thus, the competitive network learns to categorize the input vectors it sees [9].

Finding the negative distance between input vector p and the weight vectors and adding the biases b compute the net input. If all biases are zero, the maximum net input a neuron can have is 0. This occurs when the input vector p equals that neuron's weight vector. After this computation all neurons within a certain neighborhood Ni*(d) (2) of the winning neuron are updated using the Kohonen rule. Specifically, we adjust all such neurons as follows.

$$iIW1,1(q) = iIW1,1(q-1)+\alpha(p(q)-iIW1,1(q-1))$$

or

$$iw(q) = (1-\alpha)\ iw(q-1)+\alpha p(q)$$

Here the neighborhood Ni*(d) (2) contains the indices for all of the neurons that lie within a radius "d" of the winning neuron i* (1).

$$Ni(d) = \{j, dij \leq d\}$$

Thus, when a vector is presented the weights of the winning neuron and its close neighbors move toward. Consequently, after many presentations, neighboring neurons will have learned vectors similar to each other [10].

IV. Experimental Result

A SOFM classification is experimented on a IRS/1D L3 Image over a specific area for a specific period. This image of size 1200*900 consists of 4 bands 0.52 - 059 microns (B2), 0.62 - 0.68 microns (B3), 0.77 - 0.86 microns (B4) and 1.55 - 1.7 microns (B5). The aim of the classification is to distinguish between dense vegetation, coarse vegetation, water body and urban area [11]. The important requirement of the classification is the Features of the Image. Features are the characteristics of the images quantified e.g. statistical features such as mean, minimum, maximum, variance, covariance, correlation and standard deviation [12]. Other examples of the features are band ratio and difference between two images of the same area (i.e. taken at difference of time span)

Here we have shown the classified image with the various colors coding having unique color using MATLAB [14].

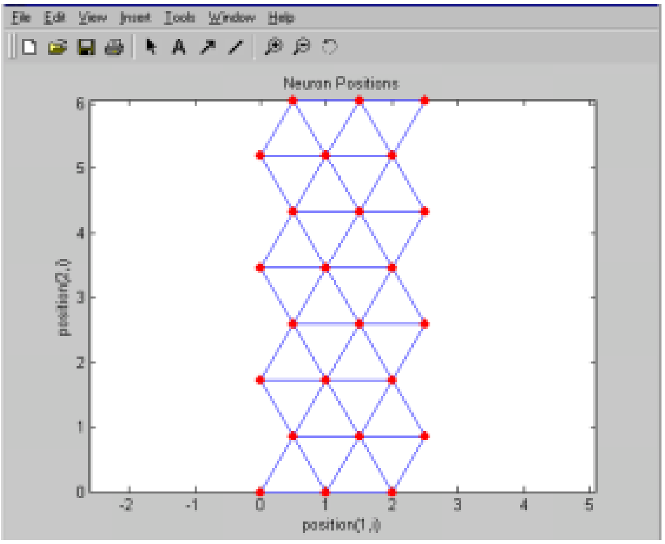


Fig.1. Neuron Positions

V. CONCLUSION

In the present work the classifications for multispectral satellite images using self organizing feature map have been done. The classification difference taken over a period of time can be used for trend analysis using SOFM. Such techniques can indeed be applied for a variety of purposes such as deforestation, archeology, urban planning and development, damage assessment, defense intelligence, and environmental monitoring, weather forecasting etc.

The work is executed using the Image Processing and Neural Network toolboxes of MATLAB because of the definite advantage of flexibility and expandability [12, 15].

REFERENCES

[1] S. Kauffman, “Origins of Order”, Oxford University Press, 1993.

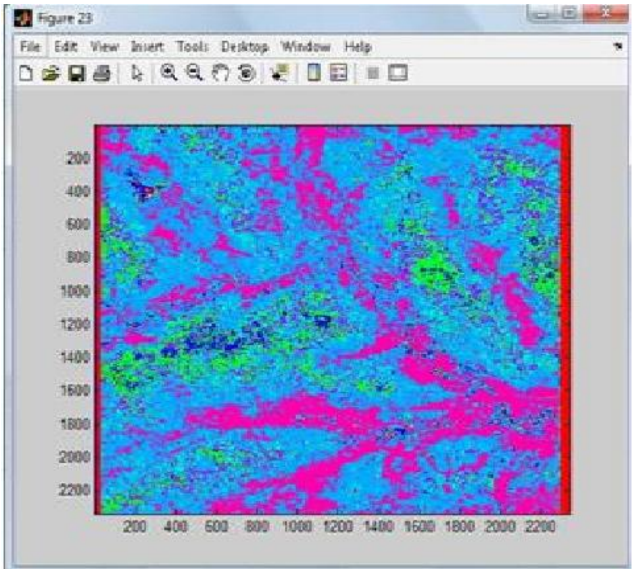


Fig.2. Classified Image

[2] B.Yam, “Dynamics of Complex Systems”, Addison-Wesley, 1997.

[3] R. Rojas, “Neural Networks: A Systematic Introduction”. Springer, Berlin, 1996.

[4] D. Rumelhart and J. McClelland, “Parallel Distributed Processing”, MIT Press, Cambridge, 1986.

[5] A. Mangal, P. Mathur and R. Govil. “Trend Analysis in satellite Imagery Using SOFM”. Apaji Institute of Mathematics & Computer Technology, Banasthali Vidhyapith, Rajasthan, India.

[6] A.S.Weigend and N.Gershenfeld, editors.Time series prediction: “forecasting the future and understanding the past”, Wesley, 1993.

[7] P. Sharma and U. Mutreja, “Analysis of Satellite Images using Artificial Neural Network”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[8] G. A. Carpenter, “The ART of adaptive pattern recognizing neural Network”, IEEE Comput. Mag., pp. 77-88, Mar. 1988.

[9] A. B. Yegnarayana, “Artificial Neural Networks”, Prentice Hall of India Pvt. Ltd, New Delhi, 1999.

[10] R. C. Gonzalez, “Richard E Woods Digital Image Processing”, Pearson Education, Inc Second Edition.

[11] H. S. Chae, S. J. Kim and J. A, Rye, “A Classification of Multi Spectral Landsat TM Data using Principal Component Analysis and ANN”, IEEE 1997.

[12] D. Singh, G. Sharma and G. Bhardwaj, “Application of Artificial Neural Network to Satellite Images Classification”, Academia Arena, vo. 4 issue 4, 2012.

[13] N. R. Euliano and J. C. Principe. “Self- Spatio Temporal Self Organising Feature Map”, Computational NeuroEngineering Laboratory, Department of Electrical Engineering, University of Florida, Gainesville, FL 32611.

[14] T.Koskela, M. Varsta, J.Heikkonen, and K. Kaski. “Time series prediction using recurrent SOM with local linear models. Int. J. of knowledge-Based Intelligent Engineering System”,Volume 2, Issue 1, pp 60-68,1998.

[15] G. J. Chappell and J. G. Taylor, “The temporal Kohonen map, Neural Networks”, Volume 6 Issue 3, pp 441-445, 1993.

Reducing Attributes in Rough Set Theory with the Viewpoint of Mining Frequent Patterns

Thanh-Trung Nguyen
Department of Computer Science
University of Information Technology,
Vietnam National University HCM City
Ho Chi Minh City, Vietnam

Phi-Khu Nguyen
Department of Computer Science
University of Information Technology,
Vietnam National University HCM City
Ho Chi Minh City, Vietnam

***Abstract*—The main objective of the Attribute Reduction problem in Rough Set Theory is to find and retain the set of attributes whose values vary most between objects in an Information System or Decision System. Besides, Mining Frequent Patterns aims finding items that the number of times they appear together in transactions exceeds a given threshold as much as possible. Therefore, the two problems have similarities. From that, an idea formed is to solve the problem of Attribute Reduction from the viewpoint and method of Mining Frequent Patterns. The main difficulty of the Attribute Reduction problem is the time consuming for execution, NP-hard. This article proposes two new algorithms for Attribute Reduction: one has linear complexity, and one has global optimum with concepts of Maximal Random Prior Set and Maximal Set.**

Keywords*—accumulating frequent patterns; attribute reduction; maximal set; maximal random prior set; mining frequent patterns; rough set

I. Introduction

Attribute reduction has played an important role in rough set applied in many fields, such as data mining, pattern recognition, machine learning. In recent years, there are many proposed reduction algorithms based on positive-region, information entropy and discernibility matrix (Qian et. al. 2011).

Attribute reduction methods have been applied by reducing inadequate features to discover hidden patterns from high dimensional data sets. Meanwhile, the nature of the original features still remained and the time consuming for pattern recognition has been decreased (Dash et. al. 2010) (Liang et. al. 2013) (Qian et. al. 2010). The characteristics of the data set are remained by keeping the important attributes. Therefore, the quality of this data set has been enhanced through the removal of redundant attributes (Sadasivam et. al. 2012). Also, rule induction can be applied in rough set theory due to attribute reduction algorithms (Yao and Zhao 2008) (Ju et. al. 2011).

One of applications of attribute reduction is gene selection. A paper presented a Quick Reduct based Genetic Algorithm (Anitha 2012) while a minimal spanning tree based on rough set theory for gene selection was introduced (Pati and Das 2013). Based on cross entropy, the relatively dispensable attributes have been omitted in the decision system and the optimal attributes set has described the same discriminative features for the original data set (Zheng and Yan 2012). In the sense of entropies, many discernibility matrixes were introduced (Wei et. al. 2013).

Based on indiscernibility and discernibility, similarities and differences of objects have been figured out and hence, attribute reduction has been carried out according to these basic theories. Attribute set is reduced by generating redacts using the indiscernibility relation of Rough Set Theory (Sengupta and Das 2012). By transforming discernibility matrix into a simplest equivalent matrix, valuable attributes have been retained while unimportant attributes will be removed from the discernibility matrix (Yao and Zhao 2009). An attribute reduction algorithm based on genetic algorithm with improved selection operator and discernibility matrix was researched and introduced (Zhenjiang et. al. 2012). Some others discussed an algorithm on discernibility matrix and Information Gain to reduce attributes (Azhagusundari and Thanamani 2013).

In addition, a proposed hybrid algorithm for large data sets was studied to overcome the shortcoming about computationally time-consuming and inefficient significance measure for more attributes with the same greatest value (Qian et. al. 2011).

Heterogeneous attribute reduction technique can be based on neighborhood rough sets by using neighborhood dependency to evaluate the discriminating capability of a subset of heterogeneous attributes. This neighborhood model reduced the attributes according to the thresholds of samples in decision positive region (Hu et. al. 2008).

In incomplete decision systems, attribute reduction methods, such as distributive reduction and positive region reduction have been given by discernibility function (Jilin et. al. 2009). To deal with these systems, a paper proposed a new attribute reduction method based on information quantity. This approach improved traditional tolerance relationship calculation methods using an extension of tolerance relationship in rough set theory (Xu et. al. 2012). Another research presented a new attribute reduction algorithm based on incomplete decision table, which improves the two aspects of time and space complexity (Yue et. al. 2012).

Handling attribute reduction problem in special systems is also a challenged issue. There are some researches in attribute reduction about dynamic data sets (Wang et. al. 2013), fuzzy sets (Chen et. al. 2012), Inconsistent Disjunctive Set-valued Ordered Decision Information System (Zhang et. al. 2012) etc.

Even, the design and implement of rough set processor in VHDL have studied on Binary Discernibility matrix and reduct calculator block (Tiwari et. al. 2012). Thereby, the speed of the operation for a dedicated hardware has been increased.

The calculation time is always a big issue in attribute reduction. A new accelerator for attribute reduction has been proposed based on perspective of objects and attributes (Liang et. al. 2013). Particle swarm optimization was a new heuristic algorithm which has been applied to many optimization problems successfully (Ding et. al. 2012). Nowadays, it is often used to solve non-deterministic polynomial (NP)-hard problem such as attribute reduction problem. Co-PSAR was introduced based on this idea to find the minimal reduction set. An algorithm based on rough set and Wasp Swarm Optimization was also introduced. It utilizes mutual information based information entropy to find core attributes, and then utilizes the significance of feature as probability information to search through the feature space for minimum attributes reduction result (Fan and Zhong 2012). A popular method in swarm intelligence is Ant Colony Optimization (ACO). A research proposed hybrid approach can help in improving classification accuracy and also in finding more robust features to improve classifier performance based on ACO (Arafat et. al. 2013).

Genetic algorithm was also researched and applied to attribute reduction. The convergence speed of algorithm is faster in global optimal solution (Zhenjiang et. al. 2012) (Liu et. al. 2013).

Besides, granular computing has been a new research approach studied to reduce the attribute in decision system (Li et. al. 2013). A paper presented a novel granularity partition model and developed a fast effective feature selection algorithm in decision systems (Sun et. al. 2012).

Some other approaches have been researched recently about Nonlinear Great Deluge Algorithm (Jaddi and Abdullah 2013), Quantization (Li et. al. 2012), attribute significance (Zhai et. al. 2012), degree of condition attributes (Qiu et. al. 2012) … They are all proved their efficiency in solving attribute reduction problem.

This article introduces an algorithm based on bit-chains and maximal random prior set. It finds out a reduction with linear time but the result is not global optimization. Therefore, another algorithm based on maximal set (a new development of maximal random prior set) and the algorithm for Accumulating Frequent Pattern (Nguyen TT and Nguyen PK 2013) to find a global optimal reduction is also proposed.

II. Formulation Model

***Definition 1* (*bit-chain*)**: $< a_1a_2 \ldots a_m >$ (for $a_i \in \{0, 1\}$) is a m-bit-chain. Zero chain is a bit-chain with each bit equals 0.

***Definition 2* (*intersection operation* $\overline{\frown}$)**: The intersection operation $\overline{\frown}$ is a dyadic operation in bit-chains space.

$< a_1a_2 \ldots a_m > \overline{\frown} < b_1b_2 \ldots b_m > = < c_1c_2 \ldots c_m >$, $a_i, b_i \in \{0, 1\}$, $c_i = \min(a_i, b_i)$

***Definition 3* (*cover operation* $\hookrightarrow$)**: A bit-chain A is said to cover a bit-chain B if and only if with every position having bit-1 turned on in B, A has a corresponding bit-1 turned on.

Let $A = < a_1a_2 \ldots a_m >$, $B = < b_1b_2 \ldots b_m >$, $(\forall b_{i=1..m} \mid (b_i = 1) \rightarrow (a_i = 1)) \Rightarrow A \hookrightarrow B$

Consequence 1: A bit-chain the result of an intersection operation and differing from zero chain is always covered by two bit-chains generating it.

$(A \overline{\frown} B = C) \wedge (C \neq 0) \Rightarrow (A \hookrightarrow C) \wedge (B \hookrightarrow C)$

***Definition 4* (*maximal random prior form* δ – S)**: The maximal random prior form of a set S of bit-chains, denoted by $\delta - S$, is a bit-chain satisfying four criteria:

- Being covered most by elements in S.
- Being covered by the first element in S.
- Having number of bit-1 turned on as much as possible.
- If there are more than one bit-chain meeting three criteria above, the bit-chain chosen to be the maximal random prior form of S is one covered by the first elements in S.

For example, consider a set of 4-bit-chains <*abcd*>:

$$S = \{ \quad \begin{array}{l} \quad a\ \ b\ \ c\ \ d \\ (\boxed{1}\ 0\ \boxed{1}\ 1); \\ (\,0\ \ 0\ \boxed{1}\ 1); \\ (\boxed{1}\ 1\ \ 0\ \ 0\,); \\ (\boxed{1}\ 0\ \boxed{1}\ 0\,) \end{array} \quad \}$$

Review three bit-chains:

<0011>: has two bit-1 turned on but is only covered by the first two bit-chains of S.

<1000>: has one bit-1 turned on and is covered by three bit-chains of S.

<0010>: has one bit-1 turned on and is covered by three bit-chains of S.

Between <1000> and <0010>, <0010> is covered by the first two elements in S, so $\delta - S$ has to be <0010>.

***Definition 5* (*maximal random prior elements*)**: Maximal random prior elements of set S of bit-chains have the following characteristics:

The first element (p_1) is form $\delta - S$

The second element (p_2) is form $\delta - S \backslash \{x \in S \mid x \hookrightarrow p_1\}$

The third element (p_3) is form $\delta - S \backslash (\{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\})$

…

The k^{th} element (p_k) is form $\delta - S \backslash (\{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\} \cup \ldots \cup \{x \in S \mid x \hookrightarrow p_{k-1}\})$

and $S = \{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\} \cup \ldots \cup \{x \in S \mid x \hookrightarrow p_k\}$

***Definition 6* (*maximal random prior set*)**: A set P containing all maximal random prior elements of a set S of bit-chains is called maximal random prior set of S.

Consequence 2: All elements in maximal random prior set *P* do not have any the same position where bit-1 turned on.

Consequence 3: When the bit-chains set is arranged in different orders, it will produce different maximal random prior sets.

Theorem 1: When the intersection operations are made between an element in *S* and elements in *P*, the results differing from zero chain will not cover each other.

Proof: According to *Consequence 2*, the results made from intersection operations of an element in *S* and elements in *P* will not have bit-1 turned on at the same position. So that, these will not cover each other.

III. Algorithm for Finding Maximal Random Prior Set

A. Idea

Consider a Boolean function *f* the intersection ($\wedge$) of *n* propositions. Each proposition in *f* is a union ($\vee$) of *m* variables a_1, a_2, ..., a_m. According to commutative law of Boolean algebra, *n* propositions of *f* can be changed into the form: $f = A_1 \wedge A_2 \wedge ... \wedge A_m$, with:

$A_1 = \wedge_{k_1} (a_1 \vee \dots)$

$A_2 = \wedge_{k_2} (a_2 \vee \dots)$ A_2 does not contain a_1

$A_3 = \wedge_{k_3} (a_3 \vee \dots)$ A_3 does not contain a_1, a_2

…

$A_m = \wedge_{k_m} (a_m \vee \dots)$ A_m does not contain $a_1, a_2, \dots, a_{m-1}$

$\forall i = 1..m; 0 \le k_i \le n \mid k_1 + k_2 + ... + k_m = n;$

$\forall k_p \ne 0; 1 \le p \le m \mid A_p = a_p \vee X_p$; X_p is a certain proposition.

So, $f = \wedge A_p = \wedge (a_p \vee X_p) = (\wedge a_p) \vee (\wedge X_p)$

Clearly, $(\wedge a_p)$ is a reduction of *f*.

If *n* propositions in *f* are transformed into a set *S* of *m*-bit-chains, the maximal random prior set *P* will be a reduction of *f*.

According to the above analysis, an algorithm is taken shape to construct maximal random prior set *P* of the bit-chains set *S* with the following main ideas:

Each element in set *S* will be inspected with the existing order in *S*. At the same time, the set *P* will be also created or modified correspondingly with the number of elements inspected in *S*.

The initial set *P* is empty. Obviously, the set *S* with one first element has the corresponding set *P* also containing only this first element.

Scanning the next element of *S*, the intersection operations ($\overline{\cap}$) made between this element and the existing elements of *P* to find out the new maximal random prior forms. If the new form is generated, it will replace the old form in *P* because this new form is covered by elements of *S* more than the old form, evidently. If the new form is not generated, obviously, the next element of *S* is one new maximal random prior form.

However, a question maybe be brought out. Whenever the next element in *S* is inspected, the elements have to carry out intersection operations with the existing elements in *P*; at that time, we have two element groups listed such as: (1) the old elements of *P*, (2) the new elements created by the intersection operations. Maybe, the new elements will cover together or cover the old elements or be covered by the old elements. Therefore, whether the set *P* is ensured the consistency as *Consequence 2* stated? The answer is “Yes” since *Consequence 1* and *Theorem 1* are generated to ensure this.

B. Proposed Algorithm

```
FIND_MaximalRandomPriorSet
Input: m-bit-chains set S
Output: maximal random prior set P
1.  P = ∅;
2.  for each s in S do
3.    flag = False;
4.    for each p in P do
5.      temp = s ⩀ p;
6.      if temp <> 0 then//temp differs
                            from zero chain
7.        replace p in P by temp;
8.        flag = True;
9.        break;
10.     end if;
11.   end for;
12.   if flag = False then
13.     P = P ∪ {s};//s becomes ending
                              element of P
14.   end if;
15. end for;
16. return P;
```

C. Accuracy of The Algorithm

Theorem 2: *FIND_MaximalRandomPriorSet* algorithm can find out the maximal random prior set *P* of a bit-chains set *S* with a given order.

Proof by Induction:

With number of elements in *S* is 1, the only element in *S* is also form $\delta - S$. According to the algorithm, the only element in *S* is inserted into *P*. Then, the only element in *P* satisfies the definition of maximal random prior set. Since, *Theorem 2* is correct when *S* has 1 element.

Assume that *Theorem 2* is correct when *S* has *k* elements. We need to prove *Theorem 2* is correct when *S* has *k* + 1 element, too.

Because *Theorem 2* is correct when *S* has *k* elements, we have the set *P* contains all maximal random prior elements of this set *S*.

When *S* has *k* + 1 elements, it means the original set *S* having *k* elements are added a new element.

According to *FIND_MaximalRandomPriorSet* algorithm, we make intersection operations between elements in current *P* and the new $(k + 1)^{th}$ element denoted s_{k+1} in *S* (line 4 and line 5):

- If the result of the intersection operation between s_{k+1} and an element p_i in P differs from zero chain (line 6), this result is form $\delta - S \backslash (\{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\} \cup ... \cup \{x \in S \mid x \hookrightarrow p_{i-1}\})$, with S has $k + 1$ elements. Replace p_i in P by this new result element (line 7). When s_{k+1}, together with p_i, create a new maximal random prior form, we terminate intersection operations between s_{k+1} and remaining elements in P (line 9).
- If all intersection operations between s_{k+1} and each element in P return zero chain, it means s_{k+1} does not cover any element in P. Thus, the element s_{k+1} is form $\delta - \{s_{k+1}\}$, then s_{k+1} is inserted into P (line 13).

In both cases, we receive the set P satisfying the properties of the maximal random prior set of S. So, *Theorem 2* is correct when S has $k + 1$ element.

In conclusion: *FIND_MaximalRandomPriorSet* algorithm can find out the maximal random prior set P of a bit-chains set S with a given order.

IV. Attribute Reduction in Rough Set Theory

The maximal random prior set P is useful in solving and reducing Boolean algebra functions. One of the most important applications of the set P is finding out a solution of attribute reduction problem in rough set theory.

A. Rough Set

In rough set theory, information system is a pair (U, A), where U is a non-empty finite set of objects and A is a non-empty finite set of attributes. A decision system is any information system of the form (U; $A \cup \{d\}$), where $d \notin A$ is decision attribute.

TABLE I. A Decision System "Play Sport"

	Wind	Temperature	Humidity	Outlook	Play Sport
x_1	Strong	Hot	Normal	Sunny	Yes
x_2	Strong	Mild	Normal	Rain	No
x_3	Weak	Hot	Normal	Rain	No
x_4	Weak	Cool	High	Rain	Yes

With $|U|$ denotes cardinal of U, discernibility matrix of a decision system is a symmetric $|U|$x$|U|$ matrix with each entry $c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\}$ if $d(x_i) \neq d(x_j)$, otherwise $c_{ij} = \varnothing$.

TABLE II. Discernibility Matrix of Decision System "Play Sport"

	x_1	x_2	x_3	x_4
x_1	∅	∅	∅	∅
x_2	b,d	∅	∅	∅
x_3	a,d	∅	∅	∅
x_4	∅	a,b,c	b,c	∅

Table II presents a discernibility matrix of decision system "Play Sport" where a, b, c, d denote Wind, Temperature, Humidity and Outlook, respectively.

Discernibility function is a Boolean function retrieved from discernibility matrix and can be defined by the formula $f = \wedge \{ \vee c_{ij} \mid c_{ij} \neq \varnothing\}$. According to *Table II*, we have discernibility function $f = (b \vee d) \wedge (a \vee d) \wedge (a \vee b \vee c) \wedge (b \vee c)$.

Discernibility function can be simplified by using laws of Boolean algebra. All constituents in the minimal disjunctive normal form of this function are all reductions of decision system (Pawlak 2003). However, simplifying discernibility function is a NP-hard problem and attribute reduction is always the key problem in rough set theory.

B. The Maximal Random Prior Set and Attribute Reduction Problem

Consider a discernibility function f retrieved from discernibility matrix of a decision system with m attributes has n constituents. Each constituent in this function will be transformed into an m-bit-chain, with each bit denotes an attribute. The function will be converted into a set S has n bit-chains. The maximal random prior set P of the set S is the simplification of discernibility function f.

Set P shows (some) reduction(s) of function f. With each bit-chain in P, the positions where bit-1 is turned on need to be noticed. Value 1 of a bit means that the corresponding attribute will appear in reduction of f. The collection of all attributes retrieved from set P is a simplification of discernibility function f.

Example: According to discernibility function f of decision system in *Table I*, the set S includes:

$$\begin{array}{ll} S = \{ & (0\;1\;0\;\boxed{1}); \\ & (1\;0\;0\;\boxed{1}); \\ & (1\;\boxed{1\;1}\;0); \\ & (0\;\boxed{1\;1}\;0) \qquad \} \end{array}$$

Initialize $P = \varnothing$. Scan all elements in S

$S[1] = (0\ 1\ 0\ 1) \rightarrow$ insert $(0\ 1\ 0\ 1)$ into $P \rightarrow P = \{ (0\ 1\ 0\ 1) \}$

$S[2] = (1\ 0\ 0\ 1) \rightarrow (1\ 0\ 0\ 1) \overline{\cap} (0\ 1\ 0\ 1) = (0\ 0\ 0\ 1) \rightarrow$ replace $(0\ 1\ 0\ 1)$ in P by $(0\ 0\ 0\ 1) \rightarrow$ P = $\{ (0\ 0\ 0\ 1) \}$

$S[3] = (1\ 1\ 1\ 0) \rightarrow (1\ 1\ 1\ 0) \overline{\cap} (0\ 0\ 0\ 1) = (0\ 0\ 0\ 0) \rightarrow$ insert $(1\ 1\ 1\ 0)$ into $P \rightarrow P = \{ (0\ 0\ 0\ 1); (1\ 1\ 1\ 0) \}$

$S[4] = (0\ 1\ 1\ 0) \rightarrow (0\ 1\ 1\ 0) \overline{\cap} (0\ 0\ 0\ 1) = (0\ 0\ 0\ 0)$; $(0\ 1\ 1\ 0) \overline{\cap} (1\ 1\ 1\ 0) = (0\ 1\ 1\ 0) \rightarrow$ replace $(1\ 1\ 1\ 0)$ in P by $(0\ 1\ 1\ 0) \rightarrow P = \{ (0\ 0\ 0\ 1); (0\ 1\ 1\ 0) \}$

$(0\ 0\ 0\ 1) \rightarrow d$ and $(0\ 1\ 1\ 0) \rightarrow b \vee c$

So, minimal function $f = d \wedge (b \vee c)$.

In conclusion, $(d \wedge b)$ and $(d \wedge c)$ are two reductions of discernibility function f.

V. Experimentation 1

FIND_MaximalRandomPriorSet algorithm is developed and tested on a personal computer with specification: Windows 7 Ultimate 32-bit, Service Pack 1 Operating System; 4096MB RAM; Intel(R) Core(TM)2 Duo, E7400, 2.80GHz; 300GB

HDD. Programming language is C#.NET on Visual Studio 2008. The results of some testing patterns:

TABLE III. SOME TESTING PATTERNS OF *FIND_MAXIMALRANDOMPRIORSET* ALGORITHM

Length of bit-chain	Number of bit-chains	Time (unit: second)
10	1,000,000	0.2184004
10	2,000,000	0.3900007
10	5,000,000	1.0140017
10	10,000,000	2.0436036
25	1,000,000	0.2808005
25	2,000,000	0.546001
25	5,000,000	1.123202
25	10,000,000	2.8236049
50	1,000,000	0.2496004
50	2,000,000	0.7176013
50	5,000,000	1.9032033
50	10,000,000	3.978007
60	1,000,000	0.3744007
60	2,000,000	0.7644014
60	5,000,000	1.9344034
60	10,000,000	4.1964073

Attribute Reduction based on bit-chains and maximal random prior set has just been introduced. It found a reduction with linear time but the result is not global optimization. The following example will show this problem clearly:

$$
\begin{array}{ll}
 & \begin{array}{cccccc} a & b & c & d & e & g \end{array} \\
S = \{ & (\begin{array}{cccccc} 1 & 1 & 0 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 1 & 1 & 0 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 0 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 1 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 1 \end{array}) \quad \}
\end{array}
$$

When applying *FIND_MaximalRandomPriorSet* algorithm to S, we have:

P = { (1 0 0 0 0 0); (0 0 1 0 0 0); (0 1 0 1 1 1) }

(1 0 0 0 0 0) → a; (0 0 1 0 0 0) → c; (0 1 0 1 1 1) → $b \vee d \vee e \vee g$

So, minimal function $f = a \wedge c \wedge (b \vee d \vee e \vee g)$. Hence, $(a \wedge c \wedge b)$, $(a \wedge c \wedge d)$, $(a \wedge c \wedge e)$, and $(a \wedge c \wedge g)$ are four reductions of discernibility function f.

But, if the order of the elements in S is changed as follows:

$$
\begin{array}{ll}
 & \begin{array}{cccccc} a & b & c & d & e & g \end{array} \\
S = \{ & (\begin{array}{cccccc} 1 & 1 & 0 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 1 & 1 & 0 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 1 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 1 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 1 \end{array}) \\
 & (\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \end{array}); \\
 & (\begin{array}{cccccc} 0 & 0 & 1 & 0 & 0 & 0 \end{array}); \quad \}
\end{array}
$$

then we have:

P = { (0 0 0 1 0 0); (1 0 0 0 0 0); (0 0 1 0 0 0) }

(0 0 0 1 0 0) → d; (1 0 0 0 0 0) → a; (0 0 1 0 0 0) → c

So, minimal function $f = d \wedge a \wedge c$. This is also the reduction of discernibility function f. Now, we can see that this result is better than the above one because it emphasize the importance of attribute d (the values of d show the difference up to 6 times between the objects), and it also is succinct.

Obviously, with an arbitrary order of elements in S, *FIND_MaximalRandomPriorSet* algorithm can not find out the best result.

The next section is going to propose a new model which is based on maximal set (a new development of maximal random prior set) and *NewRepresentative*, the algorithm for Accumulating Frequent Patterns (Nguyen TT and Nguyen PK 2013) to find a global optimal reduction.

VI. MAXIMAL SET

Definition 7 **(*maximal form* ε – S)**: The maximal form of a set S of bit-chains, denoted by $\varepsilon - S$, is a bit-chain which is covered most by elements in S.

Definition 8 **(*maximal elements*)**: Maximal elements of set S of bit-chains have the following characteristics:

The first element (q_1) is form $\varepsilon - S$

The second element (q_2) is form $\varepsilon - S \backslash \{x \in S \mid x \hookrightarrow q_1\}$

The third element (q_3) is form $\varepsilon - S \backslash \{x \in S \mid x \hookrightarrow q_1\} \cup \{x \in S \mid x \hookrightarrow q_2\}$

…

The k^{th} element (q_k) is form $\varepsilon - S \backslash \{x \in S \mid x \hookrightarrow q_1\} \cup \{x \in S \mid x \hookrightarrow q_2\} \cup \ldots \cup \{x \in S \mid x \hookrightarrow q_{k\text{-}1}\}$

and $S = \{x \in S \mid x \hookrightarrow q_1\} \cup \{x \in S \mid x \hookrightarrow q_2\} \cup \ldots \cup \{x \in S \mid x \hookrightarrow q_k\}$

Definition 9 **(*maximal set*)**: A set Q, which contains all maximal elements of a bit-chain set S, is called maximal set of S.

VII. THE ALGORITHM FOR FINDING MAXIMAL SET

In one of our previous papers, we introduced an algorithm to find out all frequent patterns of S set of transactions. Each transaction is a bit-chain with bit locations stand for a type of items. If bit-1 is turned on, it means customer bought it in transaction and otherwise. The algorithm for accumulating frequent patterns finds out the representative set P^* every time when a new bit-chain added to S (Nguyen TT and Nguyen PK 2013). The below is the full algorithm:

```
NewRepresentative
Input: $P^*$ is a representative set of S,
       z is a bit-chain added to S.
Output: The new representative set $P^*$ of
S ∪ {z}.
    M = ∅  // $M$: set of new elements of $P^*$
    flag1 = 0
    flag2 = 0
    for each x ∈ $P^*$ do
```

```
  q = x o [z; 1]
  if q ≠ 0 // q is not a chain with all bits 0
    if x ⊆ q then $P^*$ = $P^*$ \ {x}
    if [z; 1] ⊆ q then flag1 = 1
    for each y ∈ M do
      if y ⊆ q then
        M = M \ {y}
        break for
      endif
      if q ⊆ y then
        flag2 = 1
        break for
      endif
    endfor
  else
    flag2 = 1
  endif
  if flag2 = 0 then M = M ∪ {q}
  flag2 = 0
endfor
if flag1 = 0 then $P^*$ = $P^*$ ∪ {[z; 1]}
$P^*$ = $P^*$ ∪ M
return $P^*$
```

Note (Nguyen TT and Nguyen PK 2013):

- $[z; n]$ is called a pattern. z is a bit-chain and n is the frequency ($n \in \aleph$, $n \geq 0$).
- o is called intersection operation between 2 patterns. $[a_1a_2…a_m; n_1]$ o $[b_1b_2…b_m; n_2] = [c_1c_2…c_m; n_1 + n_2]$; a_i, $b_i \in \{0, 1\}$, $c_i = \min(a_i, b_i)$
- $\subseteq$ is called contained operation between 2 patterns. $[u_1; n_1] \subseteq [u_2; n_2] \Leftrightarrow (u_1 = u_2) \wedge (n_1 \leq n_2)$
- A pattern $[u; k]$ of S is called *maximal pattern* – denoted $[u; k]_{\max\to S}$ – if and only if it doesn't exist k' such that $[u; k']_{\max\to S}$ and $k' > k$.
- P^* is *representative set* of S when $P^* = \{[u; n]_{\max\to S} \mid \nexists [v; m]_{\max\to S} : (v \hookrightarrow u \text{ and } m > n)\}$. Each element in P^* is called a *representative pattern* of S.

Consequence 4: The bit-chain of the pattern which has the highest frequency in Representative Set of a set S is the maximal form of S.

From *Consequence 4*, the *Definition 8* can be modified to become the following definition.

***Definition 10* (*maximal elements*)**: Maximal elements of set S of bit-chains have the following characteristics:

The first element (q_1) is the element $\{y_0 \in P^* \mid \exists x \in S, x \hookrightarrow y_0$ and $\forall y \in P^* \mid \exists x \in S, x \hookrightarrow y \Rightarrow y_0.frequency > y.frequency\}$

The second element (q_2) is the element $\{y_0 \in P^* \mid \exists x \in S_1, x \hookrightarrow y_0$ and $\forall y \in P^* \mid \exists x \in S_1, x \hookrightarrow y \Rightarrow y_0.frequency > y.frequency\}$, here $S_1 = S \backslash \{x \in S \mid x \hookrightarrow q_1\}$

The third element (q_3) is the element $\{y_0 \in P^* \mid \exists x \in S_2, x \hookrightarrow y_0$ and $\forall y \in P^* \mid \exists x \in S_2, x \hookrightarrow y \Rightarrow y_0.frequency > y.frequency\}$, here $S_2 = S_1 \backslash \{x \in S_1 \mid x \hookrightarrow q_2\}$

…

The (k + 1)th element (q_{k+1}) is the element $\{y_0 \in P^* \mid \exists x \in S_k, x \hookrightarrow y_0$ and $\forall y \in P^* \mid \exists x \in S_k, x \hookrightarrow y \Rightarrow y_0.frequency > y.frequency\}$, here $S_k = S_{k-1} \backslash \{x \in S_{k-1} \mid x \hookrightarrow q_k\}$

After *Definition 10* is appeared, the algorithm for finding Maximal Set is created as follows:

```
FIND_MaximalSet
Input: m-bit-chains set S
       A representative set $P^*$ of S
Output: the maximal set Q
while S is not empty do
  z = GetMaximalForm($P^*$)
  Q = Q ∪ {z}
  for each x ∈ S do
    if x o z ≠ 0 then
      S ← S \ {x}
    endif
  endfor
endwhile
return Q
```

The pseudo-code of *GetMaximalForm* algorithm is shown here:

```
GetMaximalForm
Input: A representative set $P^*$
Output: the maximal form m
m = $P^*$[1];
for each x ∈ $P^*$ do
  if m.Frequency < x.Frequency then
    m = x;
  endif
endfor
for each x ∈ $P^*$ do
  if x o m ≠ 0 then
    $P^*$ = $P^*$\{x}
  endif
endfor
return m;
```

Theorem 3: *FIND_MaximalSet* algorithm can find out the maximal set Q of a bit-chains set S.

Proof: The algorithm *FIND_MaximalSet* works as follows: First we find the element q_j, then delete elements in P^* and in S covering q_j. Repeat this until the set S is empty.

In the above, if we do not delete elements in P^* covering q_j, then we can see that the q_j we find is the same as in *Definition 10*. Hence to prove the correctness of the *FIND_MaximalSet* algorithm, we need to show that when we delete elements in P^* covering q_j then we obtain the same maximal elements as defined in *Definition 10*.

We show this by induction on j.

If $j = 1$, then q_1 is determined unambiguously. We define $S_1 = S \backslash \{x \in S \mid x \hookrightarrow q_1\}$ and $P_1 = P^* \backslash \{y \in S \mid y \hookrightarrow q_1\}$.

From *Definition 10*, q_2 is determined as follows: It is the element in P^* which is covered by at least one element in S_1 and is the one with the most frequency among such. We show now that q_2 can be determined from P_1 by the same criteria.

Indeed, if q_2 is not an element in P_1, then by definition q_2 must cover q_1. Now by the choice of q_2, q_2 must be covered by one element in S_1, called that element x_2. Since $x_2 \hookrightarrow q_2$ and $q_2 \hookrightarrow q_1$, x_2 must cover q_1. But x_2 is an element in S_1, and any element in S_1 cannot cover q_1, hence we obtain a contradiction. This shows that *Theorem 3* is true for $j = 2$.

Now assume that *Theorem 3* is true for $j = j_0$. We now prove it is true for $j = j_0 + 1$. We prove this exactly like the case from $j = 1$ to $j = 2$ above. (Q.E.D.)

Similar to the Maximal Random Prior Set, one of the applications which can integrate the Maximal Set is reducing the discernibility function of rough set. Consider an example of discernibility function $f = (a \vee b \vee d) \wedge (a \vee b \vee d) \wedge a \wedge (c \vee d) \wedge (c \vee d) \wedge c \wedge (b \vee d \vee e \vee g) \wedge (b \vee d \vee e \vee g)$ with a, b, c, d, e, g are attributes in a decision system. Change f to a set of bit-chains S:

$$\begin{array}{lcccccccl} & & a & b & c & d & e & g & \\ S = \{ & (& 1 & 1 & 0 & 1 & 0 & 0 &); \\ & (& 1 & 1 & 0 & 1 & 0 & 0 &); \\ & (& 1 & 0 & 0 & 0 & 0 & 0 &); \\ & (& 0 & 0 & 1 & 1 & 0 & 0 &); \\ & (& 0 & 0 & 1 & 1 & 0 & 0 &); \\ & (& 0 & 0 & 1 & 0 & 0 & 0 &); \\ & (& 0 & 1 & 0 & 1 & 1 & 1 &); \\ & (& 0 & 1 & 0 & 1 & 1 & 1 &) \quad \} \end{array}$$

Initialize $P^* = \varnothing$. Scan all elements in S.

* S[1] = [110100; 1]: P^* is empty. Put S[1] into P^*. P^* = {[110100; 1]}

* S[2] = [110100; 1]:

S[2] o P^*[1] = [110100; 1] o [110100; 1] = [110100; 2]

// [110100; 1] ⊆ [110100; 2]

P^* = {[110100; 2]}

* S[3] = [100000; 1]:

S[3] o P^*[1] = [100000; 1] o [110100; 2] = [100000; 3]

// S[3] ⊆ [100000; 3]

P^* = {[110100; 2]; [100000; 3]}

* S[4] = [001100; 1]:

S[4] o P^*[1] = [001100; 1] o [110100; 2] = [000100; 3]

S[4] o P^*[2] = [001100; 1] o [100000; 3] = **0** (zero chain)

P^* = {[110100; 2]; [100000; 3]; [000100; 3]; [001100; 1]}

* S[5] = [001100; 1]:

S[5] o P^*[1] = [001100; 1] o [110100; 2] = [000100; 3]

S[5] o P^*[2] = [001100; 1] o [100000; 3] = **0** (zero chain)

S[5] o P^*[3] = [001100; 1] o [000100; 3] = [000100; 4]

S[5] o P^*[4] = [001100; 1] o [001100; 1] = [001100; 2]

// [000100; 3] ⊆ [000100; 4]

// [001100; 1] ⊆ [001100; 2]

P^* = {[110100; 2]; [100000; 3]; [000100; 4]; [001100; 2]}

* S[6] = [001000; 1]:

S[6] o P^*[1] = [001000; 1] o [110100; 2] = **0** (zero chain)

S[6] o P^*[2] = [001000; 1] o [100000; 3] = **0** (zero chain)

S[6] o P^*[3] = [001000; 1] o [000100; 4] = **0** (zero chain)

S[6] o P^*[4] = [001000; 1] o [001100; 2] = [001000; 3]

// S[6] ⊆ [001000; 3]

P^* = {[110100; 2]; [100000; 3]; [000100; 4]; [001100; 2]; [001000; 3]}

* S[7] = [010111; 1]:

S[7] o P^*[1] = [010111; 1] o [110100; 2] = [010100; 3]

S[7] o P^*[2] = [010111; 1] o [100000; 3] = **0** (zero chain)

S[7] o P^*[3] = [010111; 1] o [000100; 4] = [000100; 5]

S[7] o P^*[4] = [010111; 1] o [001100; 2] = [000100; 3]

S[7] o P^*[5] = [010111; 1] o [001000; 3] = **0** (zero chain)

// [000100; 3], [000100; 4] ⊆ [000100; 5]

P^* = {[110100; 2]; [100000; 3]; [001100; 2]; [001000; 3]; [010100; 3]; [000100; 5]; [010111; 1]}

* S[8] = [010111; 1]:

S[8] o P^*[1] = [010111; 1] o [110100; 2] = [010100; 3]

S[8] o P^*[2] = [010111; 1] o [100000; 3] = **0** (zero chain)

S[8] o P^*[3] = [010111; 1] o [001100; 2] = [000100; 3]

S[8] o P^*[4] = [010111; 1] o [001000; 3] = **0** (zero chain)

S[8] o P^*[5] = [010111; 1] o [010100; 3] = [010100; 4]

S[8] o P^*[6] = [010111; 1] o [000100; 5] = [000100; 6]

S[8] o P^*[7] = [010111; 1] o [010111; 1] = [010111; 2]

// [010100; 3] ⊆ [010100; 4]

// [000100; 3], [000100; 5] ⊆ [000100; 6]

// [010111; 1] ⊆ [010111; 2]

P^* = {[110100; 2]; [100000; 3]; [001100; 2]; [001000; 3]; [010100; 4]; [000100; 6]; [010111; 2]} (*Accumulated Frequent Patterns*)

* Finding maximal set:

Initialize $Q = \varnothing$.

P^*[6] = [000100; 6] has highest frequency. Remove all elements in P^* and S which cover bit-chain 000100 of P^*[6]. Put it into Q:

Q = {000100};

P^* = {[100000; 3]; [001000; 3]};

S = {100000; 001000}

Now, $P^*[1]$ = [100000; 3] and $P^*[2]$ = [001000; 3] have the same frequencies. It allows us to select one of them as the next maximal form in maximal set. Select $P^*[1]$. Remove all elements in P^* and S which cover bit-chain 100000 of $P^*[1]$. Put it into Q:

Q = {000100; 100000};

P^* = {[001000; 3]};

S = {001000}

Finally, P^* has just only one element $P^*[1]$ = [001000; 3]. Remove all elements in P^* and S which cover bit-chain 001000 of $P^*[1]$. Put it into Q:

Q = {000100; 100000; 001000};

$P^* = \varnothing$;

$S = \varnothing$

S is empty. The algorithm is terminated. Q is the maximal set of the set S.

Q = {000100; 100000; 001000};

(000100) → d; (100000) → a; and (001000) → c

So, minimal function $f = d \wedge a \wedge c$

In conclusion, ($d \wedge a \wedge c$) is a reduction of discernibility function f.

VIII. EXPERIMENTATION 2

The experiments of proposed algorithms are conducted on a machine with Pentium(R) Dual-Core CPU, E6500 @ 2.93GHz (2 CPUs), ~2.1GHz and 2048MB main memory installed. The operating system is Windows Server 2008 R2 Enterprise 64-bit (6.1, Build 7601) Service Pack 1. Programming language is C#.NET.

Data for experiments are DataFoodMart 2000 and T40I10D100K taken from http://fimi.ua.ac.be/data/ and http://www.dagira.com/2009/12/23/foodmart-2000-universe-review-part-i-introduction website, respectively.

IX. CONCLUSION AND FUTURE WORK

The result of experimentation 1 reflects the efficiency and accuracy of *FIND_MaximalRandomPriorSet* algorithm. The complexity of this algorithm is $n.2^m$ where n is the number of bit-chains in the set S and m is the length of a bit-chain. In fact, m is often unchanged, so that, 2^m can be treated as a large constant and the complexity of *FIND_MaximalRandomPriorSet* algorithm is linear.

Mining frequent patterns is applied successfully into attribute reduction problem. *FIND_MaximalSet* algorithm maybe takes much time to execute but the result reflects the global optimization.

TABLE IV. THE RESULT FOR RUNNING THE ALGORITHMS

Data	No. of Records	No. of Attributes	AFP running time (second)	No. of FP	AR running time (second)	No. of Remaining Attributes
Food Mart 2000	10,281	18	80.979	7,184	0.140	12
T40I10 D100K	1,452,990	60	74,882.424	558,193	345.774	60

[a.] AFP: Accumulating Frequent Patterns

[b.] FP: Frequent Patterns

[c.] AR: Attribute Reduction

In future, paralleling the algorithms will be a good approach to reduce the calculation time and enhance the attribute reduction result.

Besides, integrating maximal random prior set and maximal set into practical applications will help verify their accuracy more clearly.

REFERENCES

[1] K. Anitha, “Gene selection based on rough set: applications of rough set on computational biology,” International Journal of Computing Algorithm Volume 01, Issue 02, December 2012.

[2] H. Arafat, R. M. Elawady, S. Barakat, and N. M. Elrashidy, “Using rough set and ant colony optimization in feature selection,” International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 2, Issue 1, January - February 2013.

[3] B. Azhagusundari and A. S. Thanamani, “Feature selection based on information gain,” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-2, January 2013.

[4] D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu, “A novel algorithm for finding reducts with fuzzy rough sets,” IEEE Transactions On Fuzzy Systems, Vol.20, No.2, April 2012.

[5] Rajashree Dash, Rasmita Dash, and D. Mishra, “A hybridized rough-PCA approach of attribute reduction for high dimensional data set,” European Journal of Scientific Research ISSN 1450-216X Vol.44 No.1 (2010), pp.29-38.

[6] W. Ding, J. Wang, and Z. Guan, “Cooperative extended rough attribute reduction algorithm based on improved PSO,” Journal of Systems Engineering and Electronics Vol.23, No.1, February 2012, pp.160–166.

[7] H. Fan and Y. Zhong, “A rough set approach to feature selection based on wasp swarm optimization,” Journal of Computational Information Systems 8: 3 (2012) 1037–1045.

[8] Q. Hu, D. Yu, J. Liu, and C. Wu, “Neighborhood rough set based heterogeneous feature subset selection,” Elsevier, Information Sciences 178 (2008) 3577-3594.

[9] N. S. Jaddi and S. Abdullah, “Nonlinear great deluge algorithm for rough set attribute reduction,” Journal Of Information Science And Engineering 29, 49-62 (2013).

[10] Y. Jilin, Q. Keyun, and D. Weifeng, “Attribute reduction based on generalized similarity relation in incomplete decision system,” Proceedings of the 2009 International Symposium on Information Processing (ISIP’09).

[11] L. Ju, X. Wenbin, and Z. Bei, “Construction of customer classification model based on inconsistent decision table,” International Journal of e-Education, e-Business, e-Management and e-Learning, Vol.1, No.3, August 2011.

[12] B. Li, P. Tang, and T. W. S. Chow, “Quantization of rough set based attribute reduction,” A Journal of Software Engineering and Applications, 2012, 5, 117-123.

[13] D. Li, Z. Chen, and J. Yin, “A new attribute reduction recursive algorithm based on granular computing,” Journal Of Computers, Vol.8, No.3, March 2013.

[14] J. Liang, J. Mi, W. Wei, and F. Wang, “An accelerator for attribute reduction based on perspective of objects and attributes,” Knowledge-Based Systems 9 February 2013.

[15] J. Liu, F. Min, S. Liao, and W. Zhu, “Test cost constraint attribute reduction through a genetic approach,” Journal of Information & Computational Science 10: 3 (2013) 839–849.

[16] T. T. Nguyen and P. K. Nguyen, “A new viewpoint for mining frequent patterns,” International Journal of Advanced Computer Science and Application (IJACSA), Vol.4, No.3, March 2013.

[17] S. K. Pati and A. K. Das, “Constructing minimal spanning tree based on rough set theory for gene selection,” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.4, No.1, January 2013.

[18] Z. Pawlak, “Rough sets,” The Tarragona University seminar on Formal Languages and Rough Sets in August 2003.

[19] J. Qian, D. Q. Miao, Z. H. Zhang, and W. Li, “Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation,” Elsevier, International Journal of Approximate Reasoning 52 (2011) 212- 230.

[20] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, “Positive approximation: an accelerator for attribute reduction in rough set theory,” Elsevier, Artificial Intelligence 174 (2010) 597-618.

[21] T. Qiu, Y. Lin, X. Bai, “The difference degree of condition attributes and its application in the reduction of attributes,” Journal Of Computers, Vol.7, No.5, May 2012.

[22] G. S. Sadasivam, S. Sangeetha, and K. S. Priya, “Privacy preservation with attribute reduction in quantitative association rules using PSO and DSR,” Special Issue of International Journal of Computer Application (0975-8887) on Information Processing and Remote Computing – IPRC, August 2012.

[23] S. Sengupta and A. K. Das, “Single reduct generation based on relative indiscernibility of rough set theory,” International Journal on Soft Computing (IJSC) Vol.3, No.1, February 2012.

[24] L. Sun, J. Xu, J. Ren, T. Xu, and Q. Zhang, “Granularity partition-based feature selection and its application in decision systems,” Journal of Information & Computational Science 9: 12 (2012) 3487–3500.

[25] K. S. Tiwari, A. G. Kothari, and A. G. Keskar, “Reduct generation from binary discernibility matrix: an hardware approach,” International Journal of Future Computer and Communication, Vol.1, No.3, October 2012.

[26] E. Xu, Y. Yang, and Y. Ren, “A new method of attribute reduction based on information quantity in an incomplete system,” Journal Of Software, Vol.7, No.8, August 2012.

[27] Y. Yao, and Y. Zhao, “Attribute reduction in decision - theoretic rough set models,” Information Sciences, 178(17), 3356-3373, Elsevier B.V., 2008.

[28] Y. Yao and Y. Zhao, “Discernibility matrix simplification for constructing attribute reducts,” Information Sciences, Vol.179, No.5, 867-882, 2009.

[29] D. Yue, W. Jian, and Z. Xu, “Attribute reduction algorithm based on incomplete decision table,” National Conference on Information Technology and Computer Science (CITCS 2012).

[30] Y. Zhai, C. Zhou, and Y. Sun, “Approach of rule extracting based on attribute significance and decision classification,” Journal Of Computers, Vol.7, No.2, February 2012.

[31] Q. Zhang, W. Shen, and Q. Yang, “Attribute reduction in inconsistent disjunctive set-valued ordered decision information system,” Journal of Computational Information Systems 8: 12 (2012) 5161-5167.

[32] J. Zheng and R. Yan, “Attribute reduction based on cross entropy in rough set theory,” Journal of Information & Computational Science 9: 3 (2012).

[33] W. Zhengjiang, Z. Jingmin, and G. Yan, “An attribute reduction algorithm based on genetic algorithm and discernibility matrix,” Journal Of Software, Vol.7, No.11, November 2012.

[34] F. Wang, J. Liang, C. Dang, and Y. Qian, “Attribute reduction for dynamic data sets,” Journal Applied Soft Computing Volume 13, Issue 1, January 2013, Pages 676-689.

[35] W. Wei, J. Liang, J. Wang, and Y. Qian, “Decision-relative discernibility matrixes in the sense of entropies,” International Journal of General Systems 2013.

Distributed Deployment Scheme for Homogeneous Distribution of Randomly Deployed Mobile Sensor Nodes in Wireless Sensor Network

Ajay Kumar[1], Vikrant Sharma[2] and D. Prasad[3]
[1,2,3]Department of Computer Science and Engineering, M.M. University, Mullana,
Ambala, Haryana, India.

***Abstract*—One of the most active research areas in wireless sensor networks is the coverage. The efficiency of the sensor network is measured in terms of the coverage area and connectivity. Therefore these factors must be considered during the deployment. In this paper, we have presented a scheme for homogeneous distribution of randomly distributed mobile sensor nodes (MSNs) in the deployment area. The deployment area is square in shape, which is divided into number of concentric regions centered at Base Station, these regions are separated by half of the communication range and further deployment area is divided in to numbers of regular hexagons. To achieve the maximum coverage and better connectivity MSNs will set themselves at the center of the hexagon on the instruction provided by the BS which is located at one of the corner in the deployment area. The simulation results shows that the presented scheme is better than CPVF and FLOOR schemes in terms of number of MSNs required for same coverage area and average movement required by MSNs to fix themselves at the desired location and energy efficiency.**

Keywords*—Active MSNs; Desired location; Candidate location; Communication range; Sensing range etc.

I. Introduction

A wireless sensor network typically consists of a base station (BS) and a group of geographically distributed sensor nodes (SNs) [1]. The SNs are typically small wireless devices with limited computational power, radio transmission range, storage size and battery power that cooperatively perform the task of collecting relevant data and monitor its surrounding for some change or event to occur [2]. The WSNs has its own features that not only differentiate it from other wireless networks but also craft the scope of wireless applications to disaster relief, military surveillance, habitat monitoring, target tracking and in many civic, medical and security applications [3-6]. The SNs may be left unattended in any hostile environment such as battlefields, volcanoes etc., which makes it difficult or sometimes impossible to recharge or replace their batteries. Thus, efforts must be employed to remove this deficiency of WSNs. Many protocols existing in the literature minimize energy consumption on routing paths [3-6]. Even though these approaches increase energy efficiency, they do not always prolong network lifetime; if certain nodes become popular, commonly termed as "hot spots" and present on most of the forwarding paths towards sink in the network. Some of the common characteristics of WSNs that were kept in mind before developing the scheme are discussed in [2]. Some of the major issues in the node deployment strategies are: the area covered by the SNs, connectivity among them, finding the dead SNs in the deployment area, beside these the SNs are fault prone devices due to battery loss or some physical damage during their random distribution from flying BS like helicopter, aero plane etc [7-9]. To avoid such problems, the researchers tried to identify the feasible SNs deployment strategies. However, if ample amount of energy is present to provides mobility to the SNs in the network, the randomly deploy SNs can be distributed evenly in the deployment area.

One of the most active research areas in wireless sensor networks is the coverage. Coverage is usually interpreted as how well a sensor network will monitor a field of interest. Coverage can be measured in different ways depending on the applications. In addition to coverage it is important for a sensor network to maintain connectivity. Connectivity can be defined as the ability of the sensor nodes to reach the data sink. If there is no available route from a sensor node to the data sink then the data collected by that node cannot be processed. Each node has a communication range which defines the area in which another node can be located in order to receive data. This is separate from the sensing range which defines the area a node can observe. The two ranges may be equal but are often different.

There are several factors that must be considered during the deployment of sensor networks. Many of these will be dependent upon the particular application that is being addressed. The capabilities of the sensor nodes that are being used must also be considered. Most researchers focus on a single deployment model but there are papers that attempt to develop a more general algorithm that can be used in many types of deployment.

Rest of the paper is organized as follows. Section 2 summarizes the literature survey. In Section 3 System Model is presented and Section 4 consists of network model. Network setup is explained in section 5 followed by implementation of scheme in section 6. Finally simulation results and discussion of the work is presented in section 7 followed by conclusion and future work in section 8.

II. LITERATURE SURVEY

In hostile environment where reach-ability is not possible sensor nodes are deployed randomly from some flying object like helicopter, aero plane etc. In case of random deployment to provide better connectivity Mobile sensors can be used which can relocate themselves into a network.

According to the algorithm in [10], each sensor node determining the location it needs to move to in order to provide maximum coverage. The authors perform several experiments to determine how well the network covers the area and the deployment time of the algorithm. The key weakness in this algorithm is that each node must be within the sensing range of another node in order to determine the optimal location it needs to move to, if a node is not seen by any other nodes then that node cannot determine its relative location.

In [11] authors have presented the virtual force algorithm (VFA) as a practical approach for sensor deployment. The VFA algorithm uses a force-directed approach to improve the coverage provided by an initial random placement. The VFA algorithm offers a number of important advantages. These include negligible computation time and a one-time repositioning of the sensors. Moreover, the desired sensor field coverage and model parameters can be provided as inputs to the VFA algorithm, thereby ensuring flexibility. The VFA algorithm can be made more efficient if it is provided with the theoretical bounds on the number of sensors needed to achieve a given coverage threshold. Also, there is no route plan for repositioning the sensors in the VFA algorithm, where sensor collision can happen during the repositioning. Since the current target localization algorithm considers only one target in the sensor field, it is necessary to extend the presented approach to facilitate the localization of multiple objects. Another extension lies in distributed localization and querying. Extensions to non-mobile sensor nodes and situations of sensor node failures may also be considered.

In [12] authors have presented a new scheme that are not governed by these assumptions, and thus adapt to a wider range of application scenarios. The schemes are designed to maximize sensing coverage and also guarantee connectivity for a network with arbitrary sensor communication/sensing ranges or node densities, at the cost of a small moving distance. The schemes do not need any knowledge of the field layout, which can be irregular and have obstacles/holes of arbitrary shape. Scheme is an enhanced form of the traditional virtual-force-based method, which author term the Connectivity-Preserved Virtual Force (CPVF) scheme. Authors show that the localized communication, which is the reason for its simplicity, results in poor coverage in certain cases. To improve the performance further authors described a Floor based scheme which overcomes the difficulties of CPVF and, as a result, significantly outperforms it and other state-of-the-art approaches.

In [13] authors have presented a potential field based mobile sensor network deployment strategy. Two key ideas that were dealt with in [13] are (i) forming a hexagonal structure with artificial forces generated from a potential field and (ii) its hierarchical application for wider area coverage. Hexagonal formation is shown to the optimal placement for identical sensor model in terms of coverage area. Potential field based artificial force algorithm provides a simple and efficient method to deploy large number of sensors because the force is used as control input for each node without any sophisticated control algorithms. This aspect enables constructing a hierarchical structure without any additional complexity. The main weakness of this scheme is that it cannot achieve global optimization. This is a fundamental characteristic of the potential field based method. In some cases, they have undesirable formations, where a coverage holes exist in the middle of the hexagonal structure.

While focusing on the problems of coverage, existing deployment schemes largely oversimplify the conditions for network connectivity. These schemes either assume that the communication range is large enough for sensors in geometric neighborhoods to obtain location information through local communication, or assume a dense network that remains connected. In this research work we will propose a deployment scheme for mobile wireless sensor networks to support optimum coverage, while maintaining the connectivity.

III. SYSTEM MODEL

Base Station is considered to have unlimited energy with powerful transmitters (i.e. unidirectional or Omni directional) depending on the location of Base Station, which can transmit query packets within deployment area by broadcasting or multicasting. As Shown in Fig. 1; Base Station directly sends Request to specific node using its long communication range, sensor node in turn replies to the Base Station by sending a response packet using multi-hop communication (node to node communication).

The proposed model focuses on deployment of sensors nodes to achieve following results:

1) *Maximum coverage with minimum number of sensors: The d*esired location for the placement of MSNs is computed by Base Station such that there is minimum overlapping of sensing range (r_s) of adjacent MSNs.

2) *Minimizing the average movement performed by MSNs:* The randomly deployed MSNs are assigned their final positions in deployment area by Base Station in such a manner that minimum amount of movement is required. Minimizing the movement further minimizes power consumption.

3) *Minimizing the inter node communication during deployment : The m*ovement and placement of MSNs to particular location is decided and guided by Base Station, so inter node communication required for determining desired location for particular MSN is exempted.

of sensors hence making the whole system energy efficient, reliable, cost effective, and long lasting.

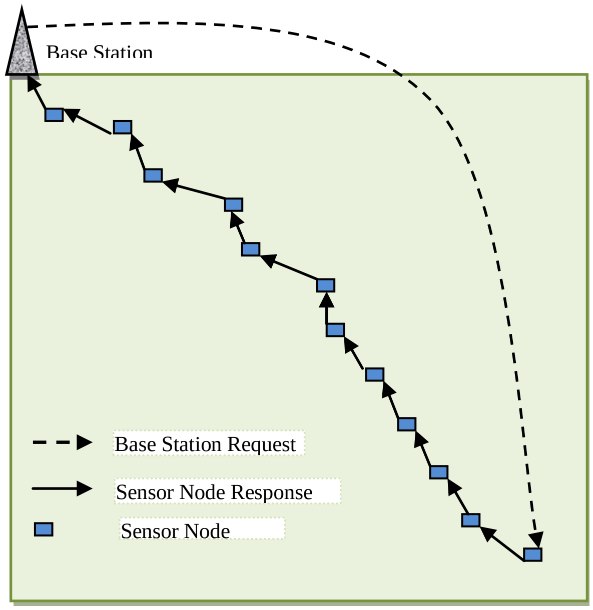


Fig. 1. Base Station Request –Response model.

4) Minimizing the energy required during deployment: All the above approaches aims at minimizing the power consumption *Minimizing the energy required during deployment:* All the above approaches aims at minimizing the power consumption

IV. Network Model

In the proposed scheme all the sensor nodes are mobile in nature. The sensor nodes are having same communication range, sensing range and computation power. The deployment area is square in shape, which is divided into number of regions these regions are separated by half of the communication range further deployment area is divided into numbers of regular hexagons. MSNs will set themselves at the center of the hexagon on the instruction provided by the BS and BS is located at one of the corner in the deployment area. We have made assumptions that all the sensor nodes fall within the deployment area when deployed randomly and the total number of SNs deployed is greater than or equal to the total number of desired locations in the deployment area. Following are the various issues in the random deployment:

A. Inefficient utilization of resources

Random deployment of sensor nodes lead to overlapping of sensing range of multiple sensors, where as some patches in deployment area remain uncovered (these patches are not in sensing range of any sensor nodes).

B. Redundant data generation

Overlapping of sensing range of multiple sensors lead to generation of packets with redundant data, which leads to congestion in the network, hence utilizes more power resources (battery).

Random deployment requires large number of sensor nodes to achieve required level of coverage which increases the system cost. So for optimum and efficient utilization of resources and for better control and management, homogenous deployment is preferred.

V. Network Setup

The sensor nodes are randomly deployed in the deployment area to be monitored by some flying object like aero plane, helicopter etc. as shown in Fig. 2.

The network setup of the presented model is divided in to the followings two phases:

A. Pre deployment Phase

Before the deployment of MSNs in the deployment area the following operations are performed at the BS.

Computation of Communication range r_c and Sensing range r_s: Let s is length of side of regular hexagon. To avoid uncovered region in the deployment area the, the sensing range (r_s) of the MSNs should be equal to s at least. The

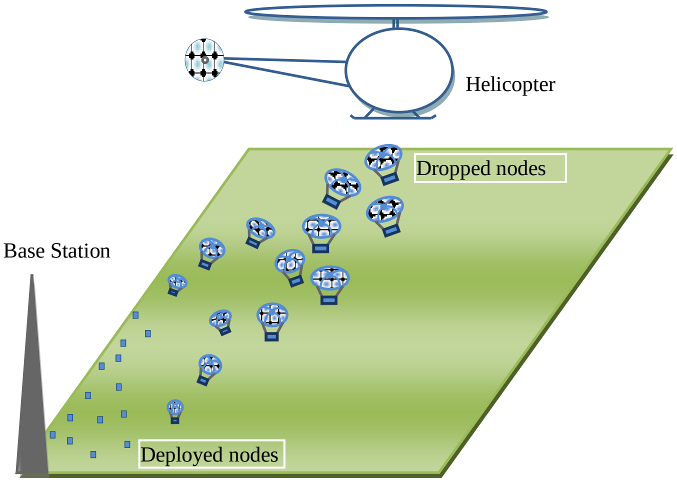


Fig. 2. Random deployment of sensor nodes using flying machine.

communication range (r_c) of the MSNs should be equal to the distance between the centers of the adjacent regular hexagon at least. The relationship between sensing range (r_s) and communication range (r_c) can be derived as follows:

Let C and D are centers of any two adjacent hexagons, AB is the common side of the above adjacent hexagon and M is the mid point of side AB as shown in the Fig. 3 By the property of regular hexagon, triangle CAB and DAB will be equilateral triangle and triangle AMC will be right angle triangle with AC as hypotenuse. By the property of right angle triangle we have the following relation:

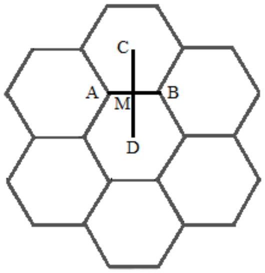


Fig. 3. Coverage pattern of the deployment area.

$$CM = \sqrt{AC^2 - AM^2} = \sqrt{s^2 - \left(s/_2\right)^2} = \sqrt{3s^2/_4}$$

Sensing range r_s in our model is equal to the length of side s and Communication r_c range will be twice of the CM i.e.

$r_c = 2*CM$; $\quad r_c = 2*CM = s*\sqrt{3} = r_s*\sqrt{3}$

1) Logical division of the deployment area: The BS divides the deployment area into n concentric regions centered at BS referred as R_1, R_2, R_3 , …….,R_n as shown in the Fig. 4 each separated by half of the communication range r_c, further entire deployment area is divided in to number of regular hexagons and center points of these hexagons are called desired locations for deployment (these are the locations where SNs will fix themselves).

2) Computation of desired locations in various regions: BS constructs n Lists RL_1, RL_2, RL_3,………,RL_n one for each region containing list of desired location in that region and add these lists to the Queue (QR) as shown in Fig. 5.

B. Post deployment Phase

The various operation performed in this phase is divided in to the following steps. In this phase MSNs are deployed in the deployment area. After the deployment MSNs find their location as in [14] and maintains at their own end. BS maintains a list called Active_Node_List[] containing Id and current location of MSNs under consideration, which are connected to BS either directly or indirectly.

The BS broadcast an ADV packet containing it own id and location in the deployment area. On receiving the ADV packet, all MSNs computes their own distance (d_{BS}) from the BS, based on d_{BS} MSNs determines the region to which they belong, as MSN belongs to region R_i if its d_{BS} lies between $(i-1) * 1/_2 * r_c$ and $i * 1/_2 * r_c$ those MSNs for which d_{BS} is less than or equal to r_c sends a RPLY packet containing their own id and location to the BS using CDMA to avoid any collision.

The maximum time taken by furthest node in i^{th} region to come in to the communication range of the placed node (i.e. node in region Front-1) is T.

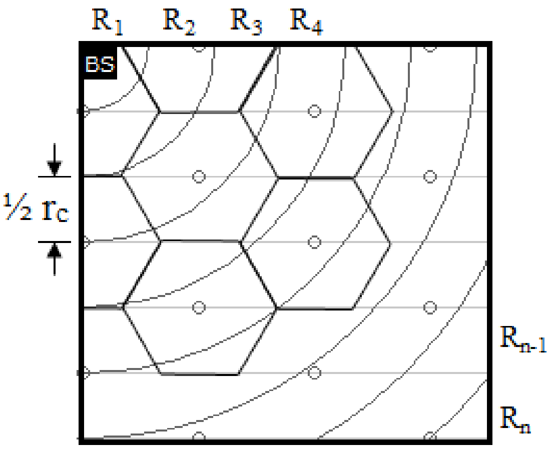


Fig. 4. Logical division of deployment area.

Algorithm 1: Homogenious distribution of MSNs

```
Step1: On receiving RPLY packet, BS updates
       Active_Node_List[] with entries of newly connected
       MSNs.
Step2: BS get all the desired locations from QR[front],
       Update Selected_Location_List[] = QR[front] and set
       i = front;
       if (Size of Active_Node_List[] = 0 ) perform Step3
       else perform Step 4.
Step 3: i=i+1;
    BS instruct the various MSNs in the next region ( i.e. $R_i$) to move towards the appropriate MSNs in the previous region (i.e. $R_{front-1}$) or BS to get connected by constructing a packet containing Id_Location_ List[] having Id and Locations of all MSNs placed in previous region (ie. $R_{front-1}$; if previous region exists) followed by its own Id and Location and send this packet for the MSNs in region $R_i$ by specifying the region in packet.

    The MSNs compare the region specified in the packet with its own region and matching MSNs compute their distance from all the locations specified in Id_Location_ List[] contained in the packet and start moving towards the MSNs with $d_{min}$ or towards BS if $d_{BS} < d_{min}$ to make themselves connected to the BS somehow and stop moving when they are within the range of placed MSNs or BS towards which they are moving, sends RPLY packet containing their Id and location to the BS. BS waits for time interval T and Perform Step1.
Step 4:
  i) BS select the appropriate MSNs (as computed by
     algorithm 2) from Active_Node_List[] to fill the various
     locations obtained in Step2, remove the corresponding
     entry from Active_Node_List[] and remove the allocated
     locations from QR[front]
     if (Size of QR[front] = 0) remove QR[front] from QR
     (ie. front = front +1);
     if(QR is empty) then exit.
```

ii)BS constructs a packet containing Id_Location_List[] having Id of selected MSNs and location computed for them, followed by Id and location of MSNs placed in previous region (i.e. $R_{front-1}$,if any) followed by its own id and location and instruct those MSNs to move to allocated location by sending the above packet to the deployment area. The MSNs in the deployment area compare their own id with the id in the list and whose id matches start moving towards the specified location, set themselves to the specified location and mark themselves as "placed" and the MSNs in the next region (ie. $R_{front+1}$) computes its distance from all the locations in the packet which are allocated to some MSNs, find the minimum distance (d_{min}) among them and start moving either_towards the MSNs with d_{min} or towards the BS if d_{BS} is less than d_{min} to make themselves connected to the BS somehow and stop moving when they are within the range of any placed MSNs or when BS comes within the range of moving MSNs and send RPLY packet containing their Id and location to the BS and then perform Step1.

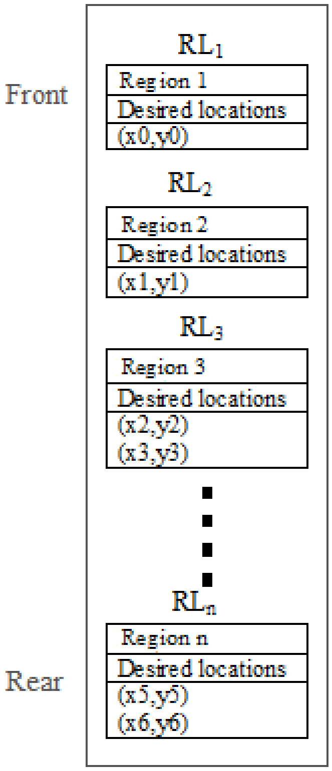


Fig. 5. Queue containing desired locations of all regions.

After all the desired locations are occupied by MSN, the BS instructs all the remaining MSNs to go to sleep mode by sending the SLEEP packet containing Id of all the remaining MSNs. These sleep MSNs may be used to occupy the uncovered spot (if any) or may be used to replace the dead MSNs in the deployment area on the instruction provided by the BS.

In Fig. 6 green arrow represents the transfer of REPLY packet containing Id and current location of newly connected MSN to Base Station using multi-hop communication, blue arrow represents transmission of packet containing Id_Location_list to selected MSNs from Active_Node_List[] (i.e. unplaced, connected MSNs from region R_0 to R_i) and to MSNs in region R_{i+1} by Base Station.

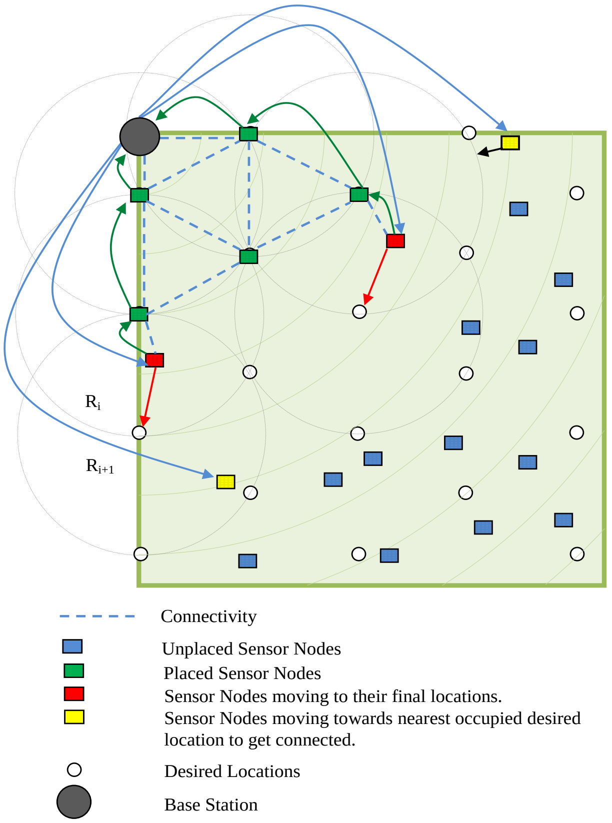

R_i
R_{i+1}
Connectivity
Unplaced Sensor Nodes
Placed Sensor Nodes
Sensor Nodes moving to their final locations.
Sensor Nodes moving towards nearest occupied desired location to get connected.
Desired Locations
Base Station

Fig. 6. Placement pattern and sequence of MSNs.

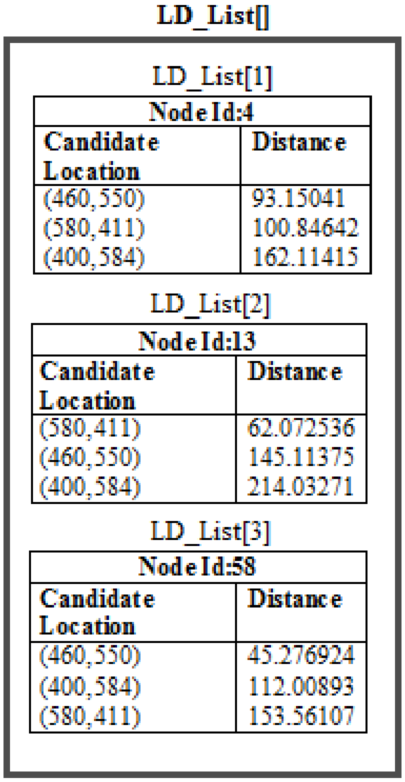
Red arrow represents the movement of selected MSNs to fix themselves to the final location specified to them in Id_Location_List[], black arrow represents the movement of MSNs in region R_{i+1} towards the nearest location in Id_Location_List[] until that location comes within communication range (r_c) of these MSNs, so that connection can be established thorugh the MSN placed at that location.

After the establishment of connection newly connected MSNs sends REPLY packet to Base Station as specified by green arrow.

Algorithm 2: Computing of the Appropriate Location for particular MSN

Step1:BS computes the distance of MSNs in Active_Node_List[] from all the locations in Selected_Location_List[] and construct tables containing candidate Location and their Distance sorted according to distance and store these tables in LD_List[] as shown in Fig. 7 using this list, BS

```
        constructs Id_Location_List[] as shown in Fig. 8
        containing MSNs Id and their Final Location.
Step2:  For i = 1 to Size_of Active_Node_List[] ; repeat Step3
        to Step5
Step3:  $Table_{min}$= first table in the list of location-distance tables
Step4:  Take first record from $Table_{min}$ and compare its location
        value with the location value of first record of all the
        other tables.
Step5:  If  location value matches
        {
          $Table_{min}$= table with minimum distance value in first
          record
          $Table_{temp}$ = other table
           If (size of $Table_{temp}$ >1)
                Remove the first record from $Table_{temp}$
           Else
                Remove $Table_{temp}$ and Go to step 3
        }
        Else
        {
          Node Id = Id of node whose table is $Table_{min}$ ;
          Final Location= location value at first record in
          $Table_{min}$ ;
          Add Node Id and Final Location to Node-Location
          Table;
          Remove $Table_{min}$ from list of location-distance tables;
        }
```



LD_List[]

LD_List[1]

Node Id:4	
Candidate Location	Distance
(460,550)	93.15041
(580,411)	100.84642
(400,584)	162.11415

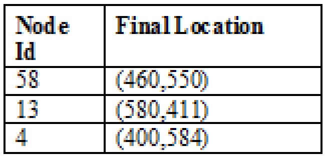
LD_List[2]

Node Id:13	
Candidate Location	Distance
(580,411)	62.072536
(460,550)	145.11375
(400,584)	214.03271

LD_List[3]

Node Id:58	
Candidate Location	Distance
(460,550)	45.276924
(400,584)	112.00893
(580,411)	153.56107

Fig. 7. List of tables containing Candidate locations and their distance from MSNs under consideration.



Node Id	Final Location
58	(460,550)
13	(580,411)
4	(400,584)

Fig. 8. List representing Id_Location_List[]

VI. Implementation

There are a number of robots that can be used as mobile SNs in this scheme, depending on the application and area of usages; like mountains, deserts, plane, etc. For simulation the presented scheme considered the Soldier UGV (SUGV). SUGV is a man- packable small robot system, weighing less than 30 lbs, used for Urban Operations environments and subterranean features to remotely investigate the obstacles, structures and the structural integrity of facilities and utilities. The SUGV system is highly mobile for dismounted forces and capable of being re-configured for other missions by adding or removing SNs, modules, mission payloads and subsystems.

VII. Simulation Results and Discussion

The presented model is simulated in JAVA. The coverage area, number of MSNs required and average movement performed by the MSNs to fix themselves to the desired location, are the various parameters we have considered for the comparison of the developed scheme with some existing schemes. The coverage pattern of the proposed scheme and the FLOOR based scheme are presented in Fig. 9 and Fig. 10 respectively. From the result obtained we can observes that the developed scheme requires 280 MSNs to achieve 100% coverage if the value of the communication range (r_c) and sensing range (r_s) are 70m and 40m respectively, as shown in Fig. 11, whereas the number of MSNs required to achieve

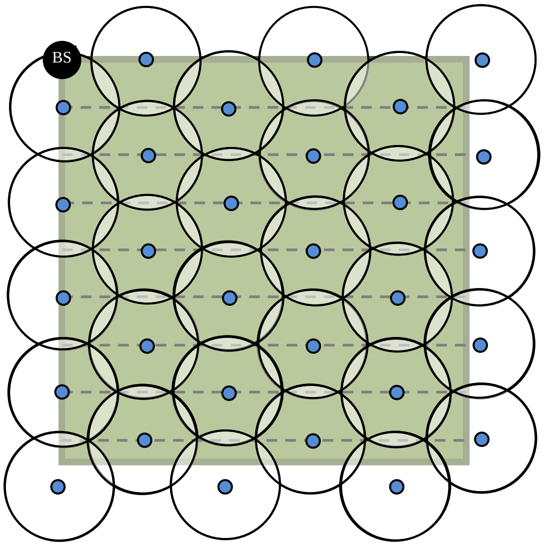


Fig. 9. 100% coverage by proposed scheme.

100% coverage is increased to 360 when the value of the communication range (r_c) and sensing range (r_s) are 60m and 35m respectively. The maximum coverage that can be achieved in FLOOR scheme is around 96% if the value of the communication range (r_c) and sensing range (r_s) are 60m and 60m respectively and requires 300 MSNs, this is due to uncovered patches in deployment area as shown in Fig. 10.

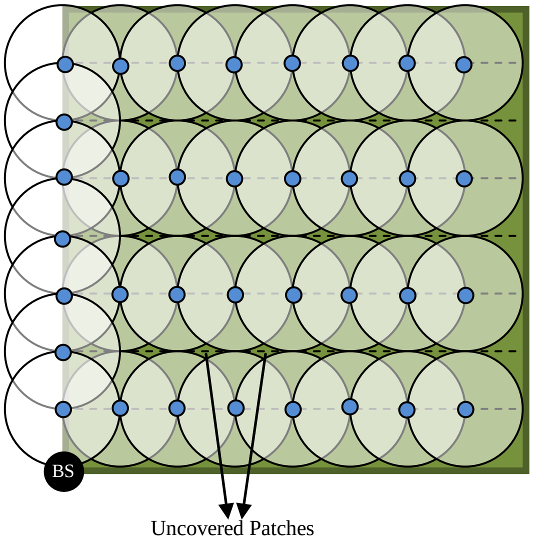


Fig. 10. Uncovered patches left in FLOOR based scheme.

Whereas the maximum coverage that can be achieved in CPVF is around 93% for the same value of the communication range (r_c), sensing range (r_s) and the number of MSNs required as in FLOOR scheme.

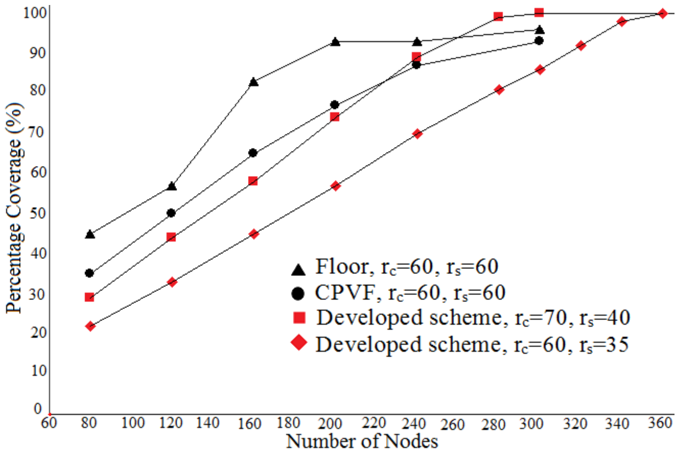


Fig. 11. Comparision of Percentage coverage versus number of Nodes

The average movement takes place by MSNs to set themselves at the appropriate location is shown in Fig. 12. We can observe that for 80 numbers of MSNs the average movement performed by the MSNs in the developed scheme is around 200m which is approximately equal to the FLOOR scheme and approximately half of the CPVF scheme. We can also observe that as the number of MSNs are increases, the average movement performed by MSNs decreases in the developed scheme whereas in FLOOR scheme and CPVF scheme as the number of MSNs increases the average movement also increases and at a value of 200 number of MSNs it becomes 330m in Floor scheme, 1060m in CPVF scheme and only 148m in the developed scheme when communication range (r_c) and sensing range (r_s) are 70m and 40 respectively and the average movement 158m for the value of communication range (r_c) and sensing range (r_s) equal to 60m and 35m respectively.

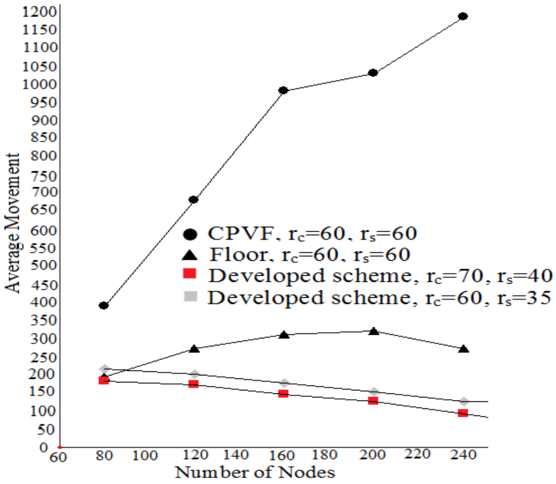


Fig. 12. Comparison of average movement versus number of nodes.

VIII. Conclusion and Future Work

In this work we have proposed a scheme for homogeneous distribution of randomly deployed MSNs to achieve maximum coverage while maintaining connectivity. The proposed scheme is energy efficient as movement of MSNs during deployment is minimized (energy consumed is directly proportional to the movement) and inter MSNs communication required for their homogenous distribution, is minimized to great extent by using Base Station to guide MSNs to set themselves to their final locations. The simulation result shows that the performance of the developed scheme is better than the earlier work.

In this scheme we have considered square deployment area and obstacle is ignored. As a future work one may think a deployment area of any irregular shape and introduce obstacle in the deployment area.

References

[1]. D. P. Agrawal and Q-A. Zeng, "*Introduction to Wireless and Mobile Systems*", Brooks/Cole publisher,2003.

[2]. I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey", Computer Networks Journal, Elsevier Science, Vol. 38(4):393-422, No. 4 pp 393–422, March 2002.

[3]. Matt Welsh, Dan Myung, Mark Gaynor, and Steve Moulton. "Resuscitation monitoring with a wireless sensor network". In Supplement to Circulation: Journal of the American Heart Association, October 2003.

[4]. G.L. Duckworth, D.C. Gilbert, and J.E. Barger. “Acoustic counter-sniper system”, In SPIE International ymposium on Enabling Technologies for Law Enforcement and Security, 1996.

[5]. Alan Mainwaring, Joseph Polastre, Robert Szewczyk, and David Culler. “Wireless sensor networks for habitat monitoring”, In First ACM International Workshop on Wireless Sensor Networks and Applications, 2002.

[6]. Robert Szewczyk, Joseph Polastre, Alan Mainwaring, and David Culler. “Lessons from a sensor network expedition”, In First European Workshop on Wireless Sensor Networks (EWSN'04), January 2004.

[7]. H. Zhang and J. C. Hou. Maintaining sensing coverage and connectivity in large sensor networks. Intl. Journal of Wireless Ad Hoc and Sensor Networks, 1(1-2):89-124, 2005.

[8]. F. Ye, G. Zhong, S. Lu, and L. Zhang, "Peas: A robust energy conserving protocol for long-lived sensornetworks," in Proceedings of the 10th IEEE International Conference on Network Protocols, Washington, DC, USA, 2002. pp. 200-201.

[9]. Siqueira, M. Fiqueiredo, A. Loureiro, J. Nogueira, and L.Ruiz, "An integrated approach for density control and routing in wireless sensor networks", in Proceedings of Parallel and Distributed Processing Symposium, Greece, April 2006, pp. 10-19.

[10]. Howard, M. J. Matari´c, and G. S. Sukhatme, “An incremental self deployment algorithm for mobile sensor networks,” Autonomous Robots, vol. 13, no. 2, pp. 113–126, Sep. 2002.

[11]. Yi Zou and Krishnendu Chakrabarty, “Sensor Deployment and Target Localization Based on Virtual Forces”, IEEE INFOCOM 2003.

[12]. Guang Tan, Member, IEEE, Stephen A. Jarvis, Member, IEEE, and Anne-Marie Kermarrec, “Connectivity-Guaranteed and Obstacle-Adaptive Deployment Schemes for Mobile Sensor Networks”, IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 8, NO. 6, JUNE 2009.

[13]. Jaeyong Lee, Avinash D. Dharne and Suhada Jayasuriya, “Potential Field Based Hierarchical Structure for Mobile Sensor Network Deployment”, Proceedings of the 2007 American Control Conference Marriott Marquis Hotel at Times Square New York City, USA, July 11-13, 2007.

[14]. Baoli Zhang and Fengqi Yu “An Energy Efficient Localization Algorithm for Wireless Sensor Networks Using a Mobile Anchor Node”, In Proceedings of the 2008 IEEE International Conference on Information and Automation June 20 -23, 2008, Zhangjiajie, China.

Towards a Fraud Prevention E-Voting System

Dr. Magdi Amer
Faculty of Computer Science
Umm A-Qura University
Makkah, KSA

Dr. Hazem El-Gendy
Faculty of Engineering
Canadian University in Egypt
Cairo, Egypt

***Abstract*—Election falsification is one of the biggest problems facing third world countries as well as developed countries with respect to cost and time. In this paper, the guidelines for building a legally binding fraud-proof Electronic-Voting are presented. Also, the limitations are discussed.**

Keywords—e-voting, security

I. Introduction

Dictators harm the countries they control. Furthermore, their harm usually extends to the rest of the world through wars they introduce and terrorist activities they host and fund. Therefore, it is the duty of everyone to work towards enforcing free election in the entire world.

Also, guaranteeing the transparency of elections, reducing the cost of elections and their times are key objectives. This motivated the trend towards electronic elections which can facilitate achieving these objectives plus being convenient. However, the move towards e-elections raises the question of guaranteeing absence of falsifications and assuring transparency.

Building electronic voting system has been an active field of research. A survey of the current legally binding electronic voting system can be found in [1], and a survey on cryptographic techniques used in e-voting can be found in [2]. There are also several research papers on increasing the reliability of the e-voting process by using verifiable voting receipts that doesn't break the privacy requirements of the election, such as [3], [4], [5].

The purpose of this paper is to provide Rules that need to be taken into consideration when building an electronic voting system to prevent vote falsification.

In this paper, the Egyptian elections will be taken as an example. The Egyptian voting process and the vote falsification practices will be explained in the next section. In the third section of the paper, software architecture guidelines will be presented. The advantages of these guidelines and the fraud prevention requirements will be discussed in the fourth section. The conclusion of this paper will be presented in the last section.

II. The Defects of the Egyptian Voting System

Egypt is one of the countries that had never conducted a proven fraud-free presidential election since it was established some thousand years ago.

The current voting system is paper based. Each voter will be assigned to a polling station based on the home address indicated in his/her National Identification card (NID), which is a unique number identifying each citizen similar to the social insurance number (SIN) used in North America. The right to vote depends on the nationality, age and criminal records. Some employees, such as police members and judges, are prevented from voting.

On the election days, voters go to their assigned polling station, sign in front of a judge that verifies their identity, take a watermarked voting ballot, choose the candidate that they want and place the ballot in the ballot box. At the end of the election, ballots are manually counted and the result will be announced.

In this manual system, voting falsification may be conducted by candidates and by voting administrators.

Candidate vote fraud can be conducted through a technique called the circulating ballot. In this technique a candidate's accomplice prints a single falsified paper ballot. The falsification doesn't need to look authentic, it only need to be good enough not to be detected when being placed in the ballot box. The candidate's accomplice goes to the polling station, takes a ballot, keeps it empty, and replaces it with the falsified ballot which he/she places in the ballot box. The candidate's accomplice will get out of the polling station and mark the empty paper ballot with the candidate he/she is representing. The candidate's accomplice will give the pre-marked ballot to the voter willing to 'sell' his/her vote for some money, items or services. The voter will go, take and empty ballot and keep it empty and replace it with the pre-marked ballot. When the voter returns the empty ballot to the candidate's agent, the voter will have earned whatever fees he/she agreed upon.

Vote falsification can also be conducted by government with the complicity of the voting authority supervising the polling stations. Voting authority may replace the ballot box with other boxes filled with ballots choosing the government's candidate. This operation is virtually undetectable as long as the fake ballot box contains the same number of ballots as the original box. This can occurs at night in case of a multi-days election, or during the transport of the ballot boxes to the counting centers.

Another vote falsification technique is to add fictitious voters to the voting database and send polling boxes corresponding to these fictitious voters to the polling station without having to replace the original ballot boxes. Another technique is to fill ballots on behalf of voters that did not show on the voting day. Both these techniques can be detected by election observers as the number of actual voters will be much lower than the number of counted votes, but it is hard to prove.

III. Rules

In this section, the system architecture for electronic voting will be presented with emphasis on the elements needed to improve the vote falsification prevention capabilities of the system.

1) Each polling station will contain an online system for voter registration that is not connected to the voting system. The purpose of the registration system is to ensure that the voter has the right to vote and that he or she did not previously vote in another polling station. This allows vote organizers to provide the voters with the option to vote in any polling station, provided that the vote organizing committee is capable of handling large numbers of voters at polling stations in more popular locations.

2) Approved voters will be given a voting token. The voting token is similar to the metallic game tokens used with game machines. These tokens need to be inserted in the voting station to allow the voter to vote and are used to prevent a voter from trying to vote multiple consecutive times at the voting stations.

3) The order of voters in the registration system should be different from the order of voters in the voting stations to prevent linking between both systems to deduce the voter's choices. The simplest way to achieve this is to have some walking distance between the registration station and the voting station that the voters will walk freely, not in lines, without maintaining a specific order.

4) Each polling station will contain one or more voting stations. A unique secret key will be assigned to each polling station and to each voting stations. Before the beginning of the election, each voting station will be programmed with the table schedule of the election days and opening and closing time. Any attempt to vote outside these pre-defined times will not be accepted. Any attempt to temper with the system time will result in the voting station locking itself and becoming unusable for the election.

5) When the voter begins the voting session, a global unique identifier will be assigned to the voter, which is called the Voter Identifier Number (VIN). The voter will choose the candidates that he or she wants from the screen and then presses a button to finalize the voting session. The choices of the voter will be printed and shown to the voter through a class, as described in [6]. The printed paper ballot will contain the VIN, the timestamp, the polling station ID, the voting station ID and a serial number representing the number of votes conducted through this voting station. The voter may press the green button to confirm the choices and terminate the voting session, in which case the paper will be dropped in the ballot box. Pressing on the red button will invalidate the voter choices and the paper will be dropped in a paper shredder, in which case the voter may use the voting station again to correct the vote.

6) Each voting station is connected to a local database. Depending on the election budget, solutions for increasing the database reliability should be implemented. In case of a complete database failure, the result will be provided by counting the printed votes in the ballot boxes.

7) At the end of each voting session, votes will be recorded on the local database. The polling station ID, the voting station ID, the VIN, the timestamp will be stored, as well as the choices of the voter, which will be encrypted using the polling station and voting station keys. The encryption algorithm chosen should combine parameters such as the VIN and the timestamp in the encryption algorithm to produce different results for each vote.

8) At the end of the election, each polling station will use its local database as well as the paper ballots to calculate the number of votes each candidate got. If both results match, the results of this polling station will be announced pending confirmation from the central voting server.

9) These votes will be sent to a central server though a network or using devices such as CDs. The center server will check that the votes were encrypted using legitimate voting station and polling station keys. Any mismatch will invalidate the electronic results of that polling station.

10) A mismatch between the paper count and the electronic count at a polling station or a mismatch between the encrypted votes and the keys of the polling station and the corresponding voting stations will invalidate the votes at that polling station. The decision on how to handle the situation will depend on the decision of the judges supervising the elections.

11) If no trace of vote manipulation is found, the central server will calculate the total votes for each candidate and publish the results. All the tables related to the vote will be made public after the decryption of the vote fields, allowing any third party to check the results.

IV. Protection Against Vote Falsification

Following the rules presented in the previous section will help achieving a high level of protection against results falsifications.

First of all, candidates will not be able to bribe voters to manipulate their votes as the candidate will have no mean of checking the choices that the voters made in the election. Nevertheless, in case of corrupted election organizers, the voter may be allowed to take a picture of the printed election paper ballot, thus providing the candidate with the proof needed to allow vote selling. This risk may be reduced by using anti-flash glass over the paper ballot and the screen of the voting station to prevent digital pictures from being taken.

Moreover, government attempts to falsify the election without hacking the programs used in the election are easy to detect and thus cannot succeed.

Replacing the actual voting boxes with fake ones is useless, as the result comparison between the paper ballot and the electronic voting database will show a mismatch which will invalidate the election at this polling station.

Entering votes for voters that did not show cannot happen. The voting station is protected from clock manipulation and from entering votes outside the pre-defined opening hours and dates of the election. The government cannot determine the list of voters that did not show till the end of the election, by which time it will be too late for the list to be useful.

Adding fictitious citizen to the voters' database and using them to enter fake vote is hard to implement. Having a person entering multiple consecutive votes during the election days will be easily detected by other voters and thus impractical. The maximum damage that a corrupted government can make is to issue multiple fake IDs to government agents and use them to vote multiple times in multiple polling stations. The logistics associated with such a plan makes this approach difficult to achieve and will place an upper limit on the number of fake votes that can be added using such a technique.

There is a risk of taking a legitimate voting station and hiding it from the public and use it to enter fake votes. This risk can be eliminated by assigning a known number of voting stations to every polling station and to make the public aware that they should report any polling station that has some missing voting stations. Moreover, statistical analysis showing a large vote bias at a specific voting station compared to other voting stations at the same polling station will be a valid proof of vote falsification.

If a corrupted government succeeds in breaking the voting application, there is hardly any solution that can be used. The work of [3], [4], [5] is very interesting as it provides a mean of detecting vote falsification while keeping the vote confidentiality, but this requires the protection of the encryption keys, which cannot be protected if the entire governmental entity supervising the election is corrupted. The only way to protect the elections in this case is to allow an international entity to supervise the election and trust it with the encryption keys. The same international entity should also be allowed to inspect voting station software during the election to detect any software manipulation.

V. Conclusion

A world with no dictatorships will be a peaceful and prosperous world. In this paper, the rules for building a system resilient against vote falsification were presented. A solution that can protect the election against a widely speed corruption in the organization supervising the election is not yet possible. The only feasible way that the current level of technology allows is to call for the establishment of an international organization that supervise the setting of the voting stations and that is trusted with the issuing and safekeeping of the needed encryption keys.

Acknowledgment

This research was conducted with the collaboration of Ambassador Mahmoud Mostafa from the Egyptian Ministry of Foreign Affairs, as part of the ministry effort to build an electronic voting system for Egyptians living abroad.

References

[1] D. Demirel, R. Frankland and M. Volkamer, "Readiness of various evoting systems for complex elections", Technische Universtität Darmstadt, Tech. Rep. TUD-CS-2011-0193, 2011, pp. 1–14.

[2] M.J. Moayed, A. Abdul Ghani and R. Mahmod, "A survey on Cryptography Algorithms in Security of Voting System Approaches", International Conference on Computational Sciences and Its Applications(ICCSA), 2008, pp. 190 - 200

[3] D. Chaum, "Secret-ballot receipts: True voter-verifiable elections", IEEE Security & Privacy, Volume: 2 , Issue: 1, 2004, pp. 38 - 47

[4] Y. Lee, S. Kim and D. Won, "How to Trust DRE Voting Machines Preserving Voter Privacy" IEEE International Conference on E-Business Engineering (ICEBE), 2008, pp. 302 – 307

[5] L. Rura, B. Issac and M. K. Haldar, "Secure Electronic Voting System Based on Image Steganography", IEEE Conference on Open Systems (ICOS), 2011, pp. 80 – 85.

[6] R. Mercuri, "A Better Ballot Box?", IEEE Spectrum, Volume: 39 , Issue: 10, 2002, pp. 46 - 50.

Impact of Medical Technology on Expansion in Healthcare Expenses

Shakir Khan, Dr. Mohamed Fahad AlAjmi

***Abstract*—the impact of medical technology on expansion in health care expenses has long been a subject of essential interest, mainly in the context of long-term outcrops of health spending, which must deal with the issue of the applicability of historical trends to future periods. The idea of this paper is to assess an approximate range for the involvement of technological alteration to growth in health spending, and to assess factors which might adjust this impact in the future. Based on the studies re-examined, we estimated that roughly half of growth in actual per capita health care costs is attributable to the beginning and diffusion of new medical technology, within an approximately probable range of 38 to 62 percent of expansion.**

Keywords—medical technology; health costs; health care; research and development

I. Introduction

Doing research on medical technology for healthcare cost enhancement has always been a great mysterious. Yet 81 percent of the primary health economists agreed with the declaration, “The primary motive to increase in the health sector’s share of gross domestic product (GDP) over the past 30 years is technological alteration in medicine”[1]. Evidently in most regions of the economy a rapid speed of technological progress is regarded as a good quality. This might not be the case for medical care where it reflects a second point of agreement.

In the past failing in medical care markets have failed to give incentives for the cost-effective condition of medical services, heartening the development and diffusion of improvement beyond the point that would overcome under spirited market conditions. Growing the role of technological change in driving growth in health spending, to the costs and benefits associated with new medical innovation reflects an acknowledgement of the long-term dilemma posed by historically unsustainable rates of growth in medical costs. It is combined with an increasing agreement that technological advance is a major factor in driving this growth. Understanding the magnitude of technology’s historical contribution to growth in costs is very important to the analysis of the future path in medical spending.

Evaluation of macroeconomic approximations, it was considered that technological modifications accounted for around fifty percent (within a “probable” range of 38 to 62 percent) of growth in real per capita health expenses restricted on statements. However, even as we package that the increase of new medical technology is the most important factor in explaining the growth but important issues was as follow; first, a primary issue surrounding the rapid growth in health care costs is not the truth of such kind development but the possibility about reflecting an inefficient use of resources that would be more valuable to the society if applied somewhere else. To what level is spending on the growth and application of new technologies defensible by the paybacks? It conveys that research is beginning to attempt to value the benefits conveyed by new technologies. Detecting whether these returns have gone beyond approximated costs and to find out where the marginal profits of new expenses are possible to be the greatest. Second, to the extent that some expenses on innovative technologies is incompetent. How can the existing enticements be changed so as to motivate a more suitable reflection of cost effectiveness? The most important force to new research on these issues is recent rapid institutional change in the delivery of health services, particularly the rise of managed care. Resulting changes in incentives surrounding the development and introduction of new technology have the potential to alter both the future direction of medical innovation and the path of growth in health spending.

Our objective in this paper is to review an estimate of a probable range for the magnitude of the historical contribution in technological change to medical spending growth based on the body of existing macroeconomic, residual-based estimates, augmenting this work possibly based on additional research.

II. Macro-Economic Evaluation: How Significant Is Technological Alter?

Health spending development has exceeded annual growth in GDP by an average of 2.2 percentage points for the period from 1940 through 1998; dynamically push the share of the economy’s resources dedicated to health sharply upwards. The search for a justification of this strong-minded tendency has a correspondingly long history. While the expansion and diffusion of new medical technology has always been identified as an aspect in health expenses growth.

Expenses on innovative medical technology take account of growth associated with the development of diffusion of medical modernisms following their original introduction, to an objective of infiltration where no further distribution occurs in the deficiency of changes on other factors. Data considerations effectively prohibit the direct quantity of technology’s role on aggregate health care expenses. Because of the approximations of the magnitude, the impact of technology health spending falls mainly into two categories. First, macro-economic estimates which relies on an indirect approach, trying to estimate the contribution of technology growth by considering the involvement of all other factors that control health spending. Second, educated guess based on analysis of the transform in behaviour patterns for an instance of patients over time which speaks to the impact of specific technologies particularly diagnoses within occurrences of care. Providing their focus on following the use of technology for residents of patients with a known diagnosis, these studies cannot confine the belongings of diffusion of new measures to broader populations. While such studies provide critical insights into the nature of technology's contribution to growth, a high degree of variability in the results across diagnosis and time period to rule out sweeping statement up to a level of agreement.

Evaluation of the involvement of technological change into cumulative growth must therefore rest primarily on studies based on the macro-economic residual approach, which provides the only comprehensive estimates of the contribution of technology to growth in the spending. However, a review of the methodology involved in the compilation of such estimates indicates that they must be applied with care. Any estimate based on the attribution of a residual after accounting for other factors will be sensitive to the identification of factors contributing to growth, as well as to the numerous assumptions necessary to evaluate the role of each factor. In addition, these estimates convey no information as to the nature of the process through which technology influences costs. One important objective of this review is to evaluate this sensitivity of residual-based estimates to the underlying assumptions, and the degree of uncertainty associated with each of the major assumptions. Based on this discussion, we produce our own estimate of the probable range for the contribution of technology to spending growth.

Ideally, macro-economic estimates of the residual growth attributable in technological change produce an estimate of the growth in health spending that would have occurred if medical technology had remained static. Suppose medical technology was frozen at a given point of time – what rate of growth in health spending would result from change in non-technology factors? Such factors include rising demand for medical services due to population growth and aging, the changing breadth and nature of health insurance coverage, rising real incomes, economy-wide inflation, and medical price inflation above economy-wide rates. Isolating the effects of technology requires that we appropriately and convincingly account for the contribution of all non-technology factors driving growth in health costs.

In calculating approximately the input of technological change by this system, our most important objective is to create a summary measure of the significance of technological change in explaining growth. However, in understanding the development through which this effect occurs, it is important to note that this contribution is dependent upon incentives inherent in financial and institutional structures within the health sector [2]. In addition to the scope that there are exchanges among the variables which power health spending growth, this methodology contains these effects in the approximate input of technological change as well. For instance, wide and more generous insurance coverage can be estimated to have significant effects on the development and dispersion of new medical technology [3].

III. Macroeconomic Residual-Based Approximations

The corrosion of health expenses growth into factors accounting for development has long been utilized as a tool to assess the comparative importance of such factors. The beginning studies include Klarman, Rice, and Cooper (1970) and Freeland and Schendler (1983) [4]. However, the centre of these earlier studies was an accounting decay of the fraction of development attributable to power on growth such as economy-wide price rises, population growth, medical inflation and population aging. The result of behavioural factors contributing to grow in requirement for medical care was not addressed. It was known that the growth detained by the residual integrated the result of many different aspects; however, no effort was made to point the residual to the technology or to any grouping of other factors. Technological transform and the growing breadth and depth of insurance treatment promoted by tax-deductibility of the employer-offered health benefits were both found to play a most important role in the constantly rising share of GDP dedicated to health spending. However, estimates of the involvement of rising insurance coverage fluctuated by a factor of ten, permitting for the diligence of a broad range of positions. [5]

The present growing agreement that technological-alter is possiblly the important factor in describing health-spending growth constantly above GDP growth has set over the past many years. Much of these trends in thought replicate enhanced information on other significant factors contributing to the development. A foremost factor contributing to this trend was the accessibility of enhanced estimates for main parameters based on the consequences of the Rand Health Insurance Experiment (HIE), a randomized investigational study of the impact of insurance treatment on health expenses and composition at the family level. [6]

A second factor contributing to increase in health spending is growing real income which was also expected to account for only a small part of growth in real health expenditure. However, the income flexibility found by the HIE was somewhat small; richer family unit consumed only very slightly more medical care. The mass of the increase in real expenditure could not be described by either of these key factors. While the study does not afford to recognize all probable non-technology factors involved in growth, Manning *et al* (1987) concluded that the contribution of these two key factors is so little as to leave the large bulk of growth in real per capita health expenditure undescribed [6]. The authors hypothesized that technological change was the possible principal factor in describing the huge residual.

Current estimates try to identify systematically all behavioral factors offering to growth and to launch a fairly accurate magnitude for each. However, as told above, for some significant suppositions; there is yet no agreement. The position taken on key issues in the health economics literature (e.g. income elasticity of demand, comparative price changes in medical services) can outcome in broadly changing approximations of residual growth. We talk about the degree of hesitation linked with each of this hypothesis.

Two current studies, Newhouse (1992) and Cutler (1995) [7], [8], try a systematic decomposition of health expenditure development, expanding their estimates to integrate at least an estimated impact for all significant behavioural factors contributing to enlarge in health spending [8] [9]. Each of these studies admits the insecurity inherent in the residual-based methodology, specified the continuous lack of clear agreement on some of important parameters. For this reason, both papers bring to close rough estimates of the magnitude of technology's contribution for growth. Newhouse comes across that "(non-technology factors) account for well under half – possibly under a quarter of the 50-years enhance in medical care spending", concluding, therefore that the residual one-half to three-quarters of growth is attributable to the preface of new technologies [8]. Cutler tried to produce a lower bounce for technology's contribution, choosing the high end of his possible range for each non-technology factor. He concluded that a minimum of half the growth in real per capita expenditure for 1940-90 can be accredited to technological change [9].

IV. Conclusion

Like innovations more commonly, procedures in medical technology help to advance our brilliance of life and provide citizens an advanced benchmark of living.

At an individual dealing level, advances in medical technology can make treatment less heavy or risky, and/or improve health outcomes, by rescheduling, reducing or getting rid of the need for further treatment. Few advances, such as going forward in e-health, improve efficiency and reduce errors. Lots of innovative programs make hospital treatment cheaper, because they minimize average length of staying life. However, because this will also free hospital beds, and because many modernizations also expand treatment frontiers, gains from any cost reductions tend to be put into additional treatments, at least in public hospitals, offsetting potential overall expenditure savings.

Moreover, some technological progresses are mainly cost increasing because they disclose completely new dealing frontiers, or because they are established with the aim of increasing patient safety or quality of care. These advances are occurring against a background of getting better overall health, but with the health enhancements being spreaded unevenly across the community. Technological advances can supply to reduce some of these inequities, such as the capability of telehealth and telemedicine to shrink the disadvantage in cost and access by knowledgeable people in rural and remote areas.

The interaction between increased levels of private health insurance with fast modernization in medical technology is possible to increase generally costs, and may increase access inequities. Current data point out that private patient treatment is around 25 per cent extra costly than municipal treatment, and that private patient expenses are rising twice as fast as public patient expenses.

Assessing the benefits and cost efficiency of technological progresses crosswise all health settings and conduct types is a most important challenge, and is the subject of significant attempt on the part of Government committees, researchers and clinicians. The different developments are not well incorporated and much greater teamwork between central and state governments is advantageous.

An incorporated national assessment development is the key to improve these processes and thus the efficiency and success of health expenses.

References

[1] Fuchs, V.R., "Economics, Values, and Health Care Reform", American Economic Review, 86:1-24, 1996.

[2] Weisbrod, B.A. "The Health Care Quadrilemma: An Essay on Technological Change, Insurance, Quality of Care, and Cost Containment, Journal of Economic Literature, June 1991, Vol. XXIX(2): 523-552.

[3] Peden, E.A. and Freeland, M.S. "An Analysis of Insurance Effects on Medical Spending: 1960-1993≅ Health Economics, 1998, 7:671-687.

[4] Klarman, H.E., Rice, D.P., Cooper, B.S. Sources of Increase in Selected Medical Care Expenditures, 1929-1969, Social Security Administration, Office of Research and Statistics, Staff paper No. 4, April 1970.

[5] Freeland, M.S. and Schendler, C.E. "National Health Expenditures: Growth in the 1980's: An Aging Population, New Technologies, and Increasing Competition≅ Health Care Financing Review, March 1983, Vol. 4(3):1-58.

[6] Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E., Leibowitz, A., and Marquis, M.S., "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment", American Economic Review, Vol. 77, No.3, June 1987.

[7] Newhouse, J.P. "Medical Care Costs: How Much Welfare Loss?" Journal of Economic Perspectives, summer 1992, 6(3):3-21.

[8] Cutler, D.M. Technology, Health Costs, and the NIH, Harvard University and the National Bureau of Economic Research. Paper prepared for the National Institutes of Health Economics Roundtable on Biomedical Research, September 1995.

[9] Cutler, D.M. Technology, Health Costs, and the NIH, Harvard University and the National Bureau of Economic Research. Paper prepared for the National Institutes of Health Economics Roundtable on Biomedical Research, September 1995.

Authors Profile

First Author was born on 5th Feb, 1978 at Kallanheri in Saharanpur district UP, India. He is working as a Researcher at College of Electronic Learning in King Saud University, Kingdom of Saudi Arabia. He received his Master of Science in Computer Science from Jamia Hamdard (Hamdard University), New Delhi, India in the year 2005 and he is PhD computer Science research scholar in Manav Bharti University, Solan (HP) India since 2010. He is member of IEEE. He has actively attended many international conferences and published various research papers in National and International conferences as well as journals. His areas of interests are in Cloud Computing, Software Engineering, Data Mining and E Learning. Apart from that he worked in the field of Software Development in different MNC companies at Noida India and expert in web development applications.

Second author was born in Kingdom of Saudi Arabia. He did Ph.D in Pharmacy from King Saud University in 2007.He chaired many position in the university and currently working as faculty member in pharmacy college affiliated to King Saud University. To date he taught nearly 280 pharmacy students, more than 30 courses. Students' level varies from primary to undergraduate levels Technology, USA and others.

Graph Mining Sub Domains and a Framework for Indexing – A Graphical Approach

K. Vivekanandan
Professor
BSMED
Bharathiar University
Coimbatore
India

A. Pankaj Moses Monickaraj
(Correspoding author)
Doctoral Scholar
Department of Computer Science
Bharathiar University
Coimbatore
India

D. Ramya Chithra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore
India

ABSTRACT: **Graphs are one of the popular models for effective representation of complex structured huge data and the similarity search for graphs has become a fundamental research problem in Graph Mining. In this paper initially, the preliminary graph related basic theorems are brushed and showcased on with various research sub domains such as Graph Classification, Graph Searching, Graph Indexing, and Graph Clustering. These are discussed with few of the most dominant algorithms in their respective sub domains. Finally a model is proposed along with various algorithms with their future projection.**

Keywords: Graph; Graph Mining; Graph Classification; Graph Searching; Graph Indexing; Graph Clustering

I. INTRODUCTION

The primary goal of data mining is to extract statistically significant and useful knowledge from data [1][2][3] which may be in any of the forms like image, text, links, vectors, tables and so on. Various forms of representing the data are available for both structured and semi-structured form. But both forms of data can be represented by a graph. Naturally this caused the vast area of research known as Graph Mining.

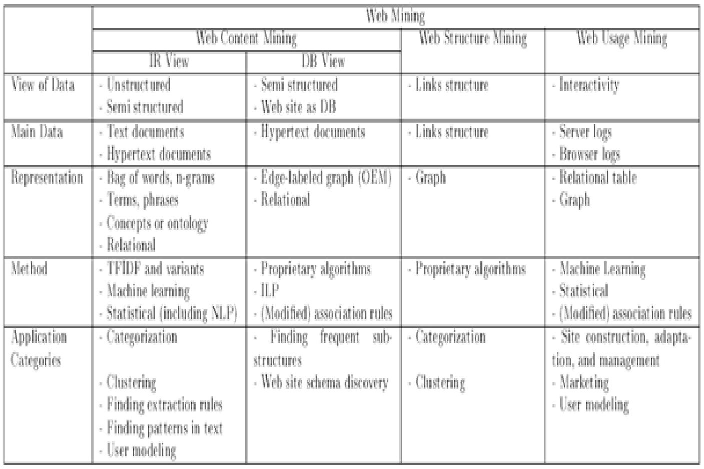
Raymond Kosala, Hendrik Blockeel in "Mining Research: A Survey", explore the connection between the web mining categories, and related agents. Interesting fact is graph structure occurs everywhere in the web mining research which is still at the budding stage [25].

From table I. , web graph is a form of representation propelled in web structure and usage mining research. In this paper, we show case the various sub domains in the field of graph mining and a model to index, update and upgrade without performance degradation.

II. RELATING GRAPH SUBSTRUCTURES WITH MATHEMATICS THEOREMS

A Graph is defined to be a set of vertexes (nodes) which are interconnected by a set of edges (links) [23].

TABLE I. Web Mining category [25]



	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View of Data	- Unstructured - Semi structured	- Semi structured - Web site as DB	- Links structure	- Interactivity
Main Data	- Text documents - Hypertext documents	- Hypertext documents	- Links structure	- Server logs - Browser logs
Representation	- Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph (OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine learning - Statistical (including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine Learning - Statistical - (Modified) association rules
Application Categories	- Categorization - Clustering - Finding extraction rules - Finding patterns in text - User modeling	- Finding frequent sub-structures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaptation, and management - Marketing - User modeling

Theorem: 1 The graph G = (V,E), where V = $\{v_1, . . . , v_n\}$ and E = $\{e_1, . . . , e_m\}$, satisfies

$$\sum_{i=1}^{n} d(v_i) = 2m$$

Corollary: Every graph has an even number of vertices of odd degree. [Figure 1]

The total sum of degree of each vertex in a graph is equal to twice the number of edges. From the number of vertices and their degrees, the number of connectivity which may be present among the vertices in the graph can be predicted which would be more useful while indexing and searching.

Theorem: 2 The vertex v is a cut vertex of the connected graph G if and only if there exist

two vertices u and w in the graph G such that (i) u ≠v, v ≠ w and u ≠ w, but (ii) v is on every u–w path.

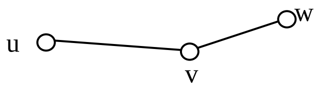


Figure 1

In this graph, u is connected to v and v is connected to w. If v is removed the connectivity is incomplete. Hence, here v is called cut vertex.

Theorem: 2 play a key role in graph classification, soon after the data are categorized according to the various conditions. The association among the content in the graph can be effectively refined by this theorem.

Theorem: 3 Every vertex of a graph G belongs to exactly one component of G. Similarly, every edge of G belongs to exactly one component of G.

Theorem:3 role comes in a graph database, when updates has to be inserted into an index, data features should be abstracted and categorized such that they can be inserted at right position in the index. Here, updates refer to the vertices and their relationship refers to the edges.

III. Glimpses Of Research Sub Domains In Graph Mining:

Using graphs as a strong method to model complex datasets, various disciplines have been recognized by various researchers in domains such as chemical [23, 24, 25], computer vision [5, 6], image and object retrieval [6, 9], and machine learning [8, 7, 9].

Enormous amount of graph data found throughout, many data mining process can be imparted but for a graph databases, it comes in different dimension. Graph classification [12], graph indexing [10][11], and graph clustering [13][18], sub graphs patterns as features are some of the major key areas of research in Graph Mining.

For example, biological structures can be stored as graphs, and in order to classify these structural graphs as active or inactive format, number of subgraph patterns are needed to build classification model [14], [15], [16].

Subgraph Isomorphism, Video Indexing, Correlated Graph Pattern Mining, Optimal Graph Pattern Mining, Approximate Graph Pattern Mining, Graph Pattern Summarization, Graph Classification, Graph Clustering, Graph Indexing, Graph Searching, Graph Kernels, Link Mining, Web Structure Mining, Work-Flow Mining, Biological Network Mining, , Improving Storage Efficiency Of Semi-Structured Databases, Efficient Indexing And Web Information Management are also some of the sub domains [23] in the field of graph mining of which few are discussed.

A. Graph Classification:

Xifeng Yan and Jiawei Han has proposed GSpan [29] (graph-based Substructure pattern mining) finds frequent substructures without candidate generation. Subgraph Mining is recursively called to grow the graphs and to find all their frequent descendants. It terminates its search when the support of a graph is less than the minimum support. It builds a new lexicographic order and maps each graph to a unique minimum Depth First Search code as its canonical label. Through this lexicographic order, it adopts the depth First search strategy to mine frequent connected sub graphs and uses a sparse adjacency list representation to store graphs.

Let {A,B,C….} be the vertices and {a,b,c….} be the connecting edges. The algorithm discovers A-^{a}A and then A-^{a}B until all frequent subgraph are discovered.

Michihiro Kuramochi and George Karyused proposed Frequent Sub Graph (FSG) [12] to find all connected subgraphs that appear frequently in a large graph database. It finds frequent subgraphs using the same level-by-level expansion adopted in Apriori [17][24].

Key features of FSG are

(1) uses a sparse graph representation minimizing both storage and computation.

(2) increases the size of frequent subgraphs by adding one edge at a time, allowing to generate the candidates efficiently

(3) uses simple algorithms of canonical labeling and graph isomorphism which work efficiently for small graphs

(4) incorporates various optimizations for candidate generation and counting which allow it to scale to large graph databases.

B. Graph Clustering:

Brian Kulis et.al has proposed a kernel approach [13] unify vector-based and graph-based approaches. The objective function for semi-supervised clustering based on Hidden Markov Random Fields, with squared Euclidean distance and a certain class of constraint penalty functions, are expressed as a special case of the weighted kernel k-means objective. It is an extension of probabilistic framework for semi supervised clustering with pairwise constraints.

This paper was based on Hidden Markov Random Fields [18]. This framework with semi-supervised clustering algorithm SS-Kernel-k means unifies vector-based and graph-based approaches using a kernel approach.

SS-Kernel-kmeans(S, k, M, C, W, tmax)

(1) Form the matrix $K = S + W$.

(2) Diagonal-shift K by adding σI to guarantee positive definiteness of K.

(3) Get initial clusters $\{\pi_c\}^k{}_c=1$ using constraints.

(4) Return $\{\pi_c{}^{(0)}\}^k{}_c=1$ = Kernel-kmeans (K, k, tmax,1, $\{\pi_c{}^{(0)}\}^k{}_{c=1}$, where 1 is the vector of all ones

C. Graph Searching:

Rosalba Giugno and Dennis Shasha has proposed an algorithm GraphGrep [20] which is an application-independent method for querying graphs, (i.e) for finding all the occurrences of a subgraph in a graph database. The interface is a regular expression graph query language Glide (a graph linear query language) the combined features from XPath and Smart acts as interface. Glide incorporates both single node and variable-length.

Steps of GraphGrep are:

(1) Build the database to represent the graphs as sets of paths

(2) Filter the database based on the submitted query to reduce the search space

(3) Perform exact matching.

The algorithm first extract all Cycle structures in a graph g, then extract all Star structures, and finally, identify the remaining structures as either Line structures or as attachments to the extracted basic structures.

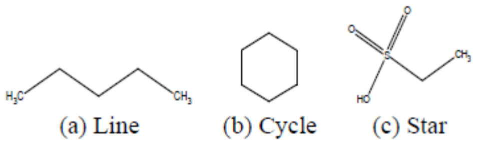


Fig. 1. Basic Structure [20]

Haoliang Jiang et.al in this paper [21] describes the transformation of a graph into a string representation, or capturing the semantics in graph data. The meaningful components in graph structures are found and are used for the most basic units in sequencing. It reduces the size of resulting sequences, but also enables semantic-based searching. Here it is approached with chemical compounds which can also be tested with protein structures as well.

D. Graph Indexing:

There are plenty of research efforts to solve the sub graph isomorphism problem for a large graph database by utilizing graph indexes of which few are listed below:

In this paper [28], Peixiang Zhao et.al proposed a new cost-effective graph indexing method based on frequent tree-features of the graph database. Effectiveness and efficiency are analyzed in three critical aspects: feature size, feature selection cost, and pruning power. To achieve better pruning, frequent tree-features (Tree),a small number of discriminative graphs (¢) are selected on demand. It has two implications: (1) the index construction by (Tree+¢) is efficient, and (2) the graph containment query processing by (Tree+¢) is efficient.

Wook Shin Han et.al has proposed iGraph [19], a framework with binary executables , heap files, B+-trees, inverted indexes, disk-based prefix trees, binary large object (BLOB) files, an LRU buffer manager, m-way posting list intersection, and external sorting.

Xifeng Yan et.al has proposed an algorithm gindex [10] which makes use of frequent substructure as the basic indexing feature.

Frequent substructures are ideal candidates as they explore the intrinsic characteristics of the data. Two techniques such as size-increasing support con straint and discriminative fragments, are introduced to reduce the size of index structure.

The design and implementation of gIndex algorithm is segmented to 5 sub sections:

(1) Discriminative fragment selection

(2) Index construction

(3) Search

(4) Verification and

(5) Incremental maintenance.

James Cheng et.al has proposed FG-index [11], novel indexing technique that constructs a nested inverted-index based on the set of Frequent subGraphs (FGs). For a graph query, FG-index returns the exact set of query answers without performing candidate verification. In case, if the query is an infrequent graph, the algorithm a candidate answer set as output which is close to the exact answer set.

The algorithm is divided into three parts:

(1) computation of T (where T is a sub graph)

(2) construction of the core FG-index,

(3) creation of Edge-index.

IV. A FRAME WORK FOR INDEXING:

Irrespective of the type of graph data, there are various mine at once algorithms to build index for any large database. After indexing, due to various updates, the index has to be restructured such that the retrieving efficiency or speed doesn't get degraded (performance). If the changes cause major performance issues, then the complete work has to be indexed from the scratch which is quite expensive and tedious.

Therefore, we propose a framework which can index with its features and update the right features at right place through search algorithms at the index.

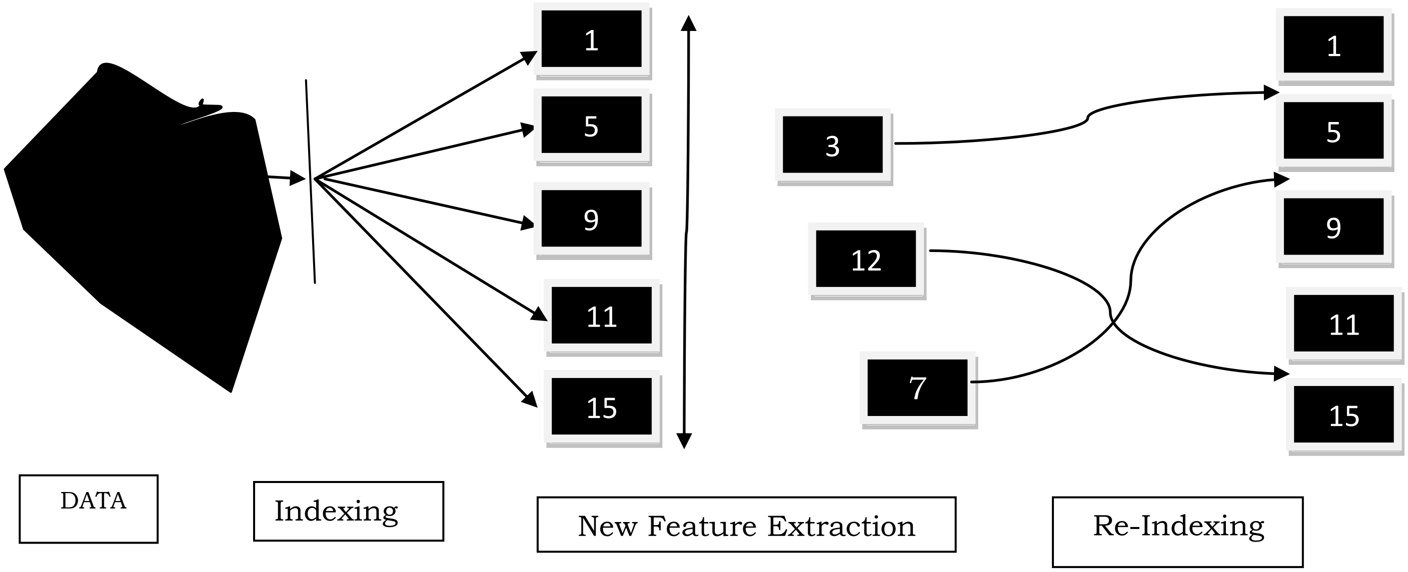


Fig. 2. ARCHITECTURAL FRAME WORK FOR GRAPH INDEXING

Mine at once indexing algorithm index any type of data. Most of the algorithms are extension or improved version of some basic techniques so a hybrid model for indexing can be built, such that indexing will be much more effective.

To upgrade the indexes with updates, the feature mining is one of the technique, in which iterative sub graph feature mining algorithm [22] is more effective in finding the upgraded parts in a graph.

Once the changes in the graph are extracted by any of the feature mining technique, right place has to be found out where the feature has to be pushed into or popped off from the index for which the basic searching techniques like BFS, DFS, G-string can be used to find the exact location where the particular extracted feature has to be pushed or popped into or off the index.

V. Conclusion

This paper includes the various areas of research fields in graph mining along with a model or architectural Framework which includes Graph Searching, Indexing and feature mining techniques. As there are plenty of mine at once algorithm, according to type of the data, effective indexing can be done by imparting the particular type of algorithm for particular data. Irrespective to the field of any applications, this model can act as a core algorithmic structure for effective indexing and upgrading the index.

References

[1] R. N. Chittimoori, L. B. Holder, and D. J. Cook. Applying the SUBDUE substructure discovery system to the chemical toxicity domain. In Proc. of the 12th International Florida AI Research Society Conference,pages 90–94, 1999

[2] A. Srinivasan, R. D. King, S. Muggleton, and M. J. E. Sternberg. Carcinogenesis predictions using ILP. In S. Dˇzeroski and N. Lavraˇc, editors, Proc. of the 7th International Workshop on Inductive Logic Programming, volume 1297, pages 273–287. Springer-Verlag, Berlin, 1997.

[3] L. Dehaspe, H. Toivonen, and R. D. King. Finding frequent substructures in chemical compounds, Proc. of the 4th International Conference on Knowledge Discovery and Data Mining, pages 30–36. AAAI Press, 1998

[4] A. Srinivasan, R. D. King, S. H. Muggleton, and M. Sternberg. The predictive toxicology evaluation challenge. In Proc. of the 15th International Joint Conference on Artificial Intelligence (IJCAI), pages 1–6. Morgan-Kaufmann, 1997.

[5] H. K¨alvi¨ainen and E. Oja. Comparisons of attributed graph matching algorithms for computer vision. In Proc. of STEP-90, Finnish Artificial Intelligence Symposium, pages 354–368, Oulu, Finland, June 1990.

[6] D. A. L. Piriyakumar and P. Levi. An efficient A* based algorithm for optimal graph matching applied to computer vision. In GRWSIA-98, Munich, 1998.

[7] C.-W. K. Chen and D. Y. Y. Yun. Unifying graph-matching problem with a practical solution. In Proceedings of International Conference on Systems, Signals, Control, Computers, September 1998.

[8] L. Holder, D. Cook, and S. Djoko. Substructure discovery in the SUBDUE system. In Proc. of the Workshop on Knowledge Discovery in Databases, pages 169–180, 1994.

[9] K. Yoshida and H. Motoda. CLIP: Concept learning from inference patterns. Artificial Intelligence, 75(1):63–92, 1995.

[10] X. Van, P. S. Yu, and J. Han. "Graph indexing: a frequent structure-based approach," In Proc. of the ACM SIGMOD international conference on Management of data, pages 335-346, 2004.

[11] J. Cheng, Y. Ke, W. Ng, and A. Lu. "Fg-index: towards verification-free query processing on graph databases," In Proc. Of the ACM SIGMOD international conference on Management of data, pp. 857-872,2007.

[12] M. Deshpande, M. Kuramochi, and G. Karypis. "Frequent substructure discovery,"Proc. 3rd IEEE Int'l Conf. Data Mining (ICDM '02), 2001.

[13] Brian Kulis, Sugato Basu, Indeljit Dhillon and Raymond Mooney "Semi-supervised graph clustering: a kernel approach " in: . Proc. Proceedings of the 22nd international conference on Machine learning ICML '05

[14] C.-W. K. Chen and D. Y. Y. Yun. Unifying graph-matching problem with a practical solution. In Proceedings of International Conference on Systems, Signals, Control, Computers, September 1998.

[15] R. N. Chittimoori, L. B. Holder, and D. J. Cook. Applying the SUBDUE ubstructure discovery system to the chemical toxicity domain. In Proc. of the 12th International Florida AI Research Society Conference, pages 90–94, 1999.

[16] V. A. Cicirello. Intelligent retrieval of solid models. Master's thesis, Drexel University, Philadelphia, PA, 1999.

[17] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, Proc. of the 20th Int. Conf. on Very Large Databases (VLDB), pages 487–499. Morgan Kaufmann, September 1994

[18] Basu, S., Bilenko, M., & Mooney, R. (2004). A probabilistic framework for semi-supervised clustering. Proc. 10th Intl. Conf. on Knowledge Discovery and Data Mining.

[19] Wook Shin Han, Jinsoo Lee, Minh Duc Pham, Jeffrey Xu Yu "iGraph: A Framework for Comparisons of Disk Based Graph Indexing Techniques", The 36th International Conference on Very Large Data Bases, September13-17,2010, Singapore. Proceedings of the VLDB Endowment, Vol. 3,

[20] Rosalba Giugno, Dennis Shasha, GraphGrep: A Fast and Universal Method for Querying Graphs, Pattern Recognition, 2002.

[21] Haoliang Jiang, Haixun Wang, Philip S. Yu, Shuigeng Zhou, GString: A Novel Approach for Efficient Search in Graph Databases, IEEE Pattern Recognition, 2002.

[22] Dayu Yuan, Prasenjit Mitra , Huiwen Yu, C. Lee Giles Iterative Graph Feature Mining for Graph Indexing, 2012 IEEE 28th International Conference on Data Engineering.

[23] Chuntao Jiang, Frans Coenen and Michele Zito, A Survey of Frequent Subgraph Mining Algorithms, The Knowledge Engineering Review, Vol. 00:0, 1–31.c 2004, Cambridge University Press

[24] Chen, M.S., Han, J. and Yu, P.S. 1996. Data Mining: An Overview from Database Perspective, IEEE Transaction on Knowledge and Data Engineering 8, 866–883.

[25] Raymond Kosala, Hendrik Block, Web Mining Research: A Survey, SGIKDD, Explorations, ACM, 2000.

[26] Akihiro Inokuchi, Takashi Washio and Hiroshi Motoda, An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data Currently being in Tokyo Research Institute, IBM, 1623-14 Shimotsuruma, Yamatoshi, Kanagawa, 242-8502, Japan.

[27] [27] Huan, 1., Wang, W., Prins, 1. Efficient mining of frequent subgraphs in the presence of isomorphism. In: Proceedings of the 3rd IEEE IntI. Conf. on Data Mining ICDM, (2003) 549-552

[28] Peixiang Zhao, Jeffrey Xu Yu, Philip S. Yu , Graph Indexing: Tree + Delta >= Graph, ACM. VLDB '07, September 2328, 2007, Vienna, Austria.

[29] Xifeng Yan Jiawei Han, gSpan: Graph-Based Substructure Pattern Mining, ICDM 2003. Proceedings,2002.

Feedback Optimal Control of Low-thrust Orbit Transfer in Central Gravity Field

Ashraf H. Owis[†], [‡]

[†]Department of Astronomy, Space and Meteorology, Faculty of Science, Cairo University

[‡]Department of Aerospace Science and Technology, Politecnico di Milano

I. Abstract

Low-thrust trajectories with variable radial thrust is studied in this paper. The problem is tackled by solving the Hamilton-Jacobi-Bellman equation via State Dependent Riccati Equation(STDE) technique devised for nonlinear systems. Instead of solving the two-point boundary value problem in which the classical optimal control is stated, this technique allows us to derive closed-loop solutions. The idea of the work consists in factorizing the original nonlinear dynamical system into a quasi-linear state dependent system of ordinary differential equations. The generating function technique is then applied to this new dynamical system, the feedback optimal control is solved. We circumvent in this way the problem of expanding the vector field and truncating higher-order terms because no remainders are lost in the undertaken approach. This technique can be applied to any planet-to-planet transfer; it has been applied here to the Earth-Mars low-thrust transfer.

II. Introduction

Historically, the optimal low-thrust transfers have been tackled first with indirect and then with direct methods. The former stem from the Pontryagin's maximum principle that uses the calculus of variations [1, 2]; the latter aim at solving the problem via a standard nonlinear programming procedure [3]. Even if it can be demonstrated that both approaches lead to the same result [4], the direct and indirect methods have different advantages and drawbacks, but they require in any case the solution of a complex set of equations: the Euler-Lagrange equations (indirect methods) and the Karush-Kuhn-Tucker equations (direct methods). The guidance designed with these methods is obtained in an open-loop context. In other words, the optimal path, even if minimizing the prescribed performance index, is not able to respond to any perturbation that could alter the state of the spacecraft. Furthermore, if the initial conditions are slightly varied (e.g. the launch date changes), the optimal solution needs to be recomputed again. The outcome of the classical problem is in fact a guidance law. expressed as a function of the time,$u = u(t), t \in [t0, tf]$, being t0 and tf the initial and final time, and u the control vector, respectively. This paper deals with the optimal feedback control problem applied to the low thrust interplanetary trajectory design. With this approach the solutions that minimize the performance index are also functions of the generic initial state x0; the outcome is in fact a guidance law written as $u = u(x0, t0, t)$, t $t \in [t0, tf]$. This represents a closed-loop solution: given the initial conditions (t0,x0) it is possible to extract the optimal control law that solves the optimal control problem. Moreover, if for any reason the state is perturbed and assumes the new value $(t0', x0') = (x0 + \delta x, t0 + \delta t)$, we are able to compute the new optimal solution by simply evaluating so avoiding the solution of another optimal control problem. This property holds by virtue of the closed loop characteristics of the control law that can be viewed as a one-parameter family of solutions. Due to such property, a trajectory designed in this way has the property to respond to perturbations acting during the transfer that continuously alter the state of the spacecraft. Another important aspect of this approach is the possibility to have robust nominal solutions. Indeed, the optimal feedback control can be analyzed and the control laws being less sensitive to changes in the initial condition can be chosen as nominal solutions. These solutions are said to be robust with respect to the initial conditions. The optimal feedback control for linear systems with quadratic objective functions is addressed through the matrix Riccati equation: this is a matrix differential equation that can be integrated backward in time to yield the initial value of the Lagrange multipliers [2]. The same problem has been tackled in an elegant fashion using the Hamiltonian dynamics and exploiting the properties of the generating functions [5]. With this approach it is possible to devise suitable canonical transformations, satisfying the Hamilton-Jacobi equation, that also verify Hamilton-Jacobi-Bellman equation of the optimal feedback control problem. The generating function technique has been extended to non-linear dynamical systems supplemented by quadratic objective functions: in this case the vector field is expanded in Taylor series and the optimal control is derived as a polynomial [6]. Nevertheless, the resulting optimal control differs from the one obtained through application of the Pontryagin principle to the nonlinear system since, in the process of series expansion and truncation, the dynamics associated to the high-order terms is neglected. Recently, the nonlinear feedback control of low-thrust orbital transfers has been faced using continuous orbital elements feedback and Lyapunov functions [7]. The analytical low-thrust optimal feedback control problem is solved, with modulated inverse-square-distance , in the frame of a nonlinear vector field, the two-body dynamics, supported by a nonlinear objective function by applying a globally diffeomorphic linearizing transformation that rearranges the original problem into a linear system of ordinary differential equations and a quadratic objective function written in a new set of variables [8].

In this work we consider the nonlinear feedback optimal control of the motion of a spacecraft under the influence of the gravitational attraction of a central body, the Sun in our case, and we would like to transfer the spacecraft from Earth to Mars. Both orbits of Earth and Mars around the Sun are assumed to be circular and coplanar. We use both radial and tangential thrust control. The nonlinear dynamics of the system will be factorized in such a way that the new factorized system is accessible. The problem is tackled by solving the State Dependent Riccati Equation (SDRE). The method is applied to Earth-Mars transfer.

III. Statement of the Problem

The equations of motion are written in polar coordinates (r, θ), in the inertial Sun-Centered frame. In order to transfer the spacecraft from Earth to Mars two components of the thrust control are used.The tangential component T_θ, and the radial component T_r.
The equations of motion are:

$$\ddot{r} - r\dot{\theta}^2 = T_r - \frac{\mu}{r^2}, \qquad r\ddot{\theta} + 2\dot{r}\dot{\theta} = T_\theta \tag{1}$$

where μ is the gravitational constant of the Sun$(1.3271 \times 10^{20} m^3/s^2)$

So as to use dimensionless variables, we take the unit of distance to be the radius of the circular orbit of the Earth around the Sun, the velocity unit is the velocity of the Earth in its orbit, the frequency is $\omega = \frac{2\pi}{T}$ where $T = 365.25$ days is the periodic time and the unit of time is $\frac{1}{\omega} = 58.131343$ days. In this system of units the gravitational constant μ is unity, and equations (1) are rewritten as:

$$\ddot{r} - r\dot{\theta}^2 = T_r - \frac{1}{r^2}, \qquad \ddot{\theta} + 2\frac{\dot{r}\dot{\theta}}{r} = \frac{T_\theta}{r} \tag{2}$$

IV. Equations of motion in state variable form

Equations (2) are then written in state variable form. The state vector $\mathbf{x}$ is chosen to be:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} r \\ \theta \\ \dot{r} \\ \dot{\theta} \end{bmatrix} \tag{3}$$

and the control vector is :

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} T_r \\ T_\theta \end{bmatrix} \tag{4}$$

Then Equation (2) can be written in the form :

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{B}(\mathbf{x})\mathbf{u} \tag{5}$$

Choosing a suitable factorization equation (5) is rewritten in the factored state variable form :

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x})\mathbf{x} + \mathbf{B}(\mathbf{x})\mathbf{u} \tag{6}$$

where :

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ x_4^2 & -\frac{1}{x_1^2 x_2} & 0 & 0 \\ 0 & 0 & -\frac{2x_4}{x_1} & 0 \end{bmatrix} \tag{7}$$

$$\mathbf{B}(\mathbf{x}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{x_1} \end{bmatrix} \tag{8}$$

V. Weak Controllability and Factored Controllability

Weak Controllability (Accessibility) and Factored Controllability

Weak Controllability (Accessibility) In order for the solution to exist, the system must be weakly controllable (Accessible). A sufficient condition for the system to be weakly controllable on $S \subset \mathbb{R}^n$ is that $\forall \mathbf{x} \in S \subset \mathbb{R}^n : rank[\Delta_c] = n$

An algorithm is given below for generating Δ_c:

1) Let $\Delta_0 = span(B) = span(b_i)$
2) Let $\Delta_1 = \Delta_0 + [a, b_i] + [b_j, b_i]$
3) Let $\Delta_k = \Delta_{k-1} + [a, d_j] + [b_i, d_j]$
4) Terminate when $\Delta_{k+1} = \Delta_k$.

Where d_j is the basis of Δ_{k-1} and the bracket $[f, g]$ is the Lie bracket of f and g

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

Applying this algorithm to the system (5) to determine its weak controllability we find that:

$$\Delta_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{x_1} \end{bmatrix}$$

whose span is 2 and

$$\Delta_1 = \Delta_2 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & \frac{-1}{x_1} \\ 1 & 0 & 0 & -2x_4 \\ 0 & \frac{1}{x_1} & \frac{2x_4}{x_1} & \frac{x_3}{x_1^2} \end{bmatrix}$$

with

$$rank[\Delta_2] = 4 = n$$

$$\forall x_1 \neq 0$$

In our problem $x_1 \geq 1$. That is the system is locally controllable.

Factored Controllability For the factored system (6) the controllability is established by verifying that the controllability matrix

$$\mathbf{M}_{cl} = [\mathbf{B}\ \mathbf{AB}\ \mathbf{A}^2\mathbf{B}\ \mathbf{A}^3\mathbf{B}]$$

has a rank equals to $n = 4\ \forall x$ in the domain.
The controllability matrix $\mathbf{M}_{cl}$ for the System (6) is:

$$\mathbf{M}_{cl} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{x_1} \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{x_1} & -\frac{2x_4}{x_1} & 0 \end{bmatrix}$$

which has a rank 4.

VI. State Dependent Riccati Equation

State Dependent Riccati Equation Consider the consider the State Dependent Linear Quadratic Regulator written as follows:

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x})\mathbf{x}(t) + \mathbf{B}(\mathbf{x})\mathbf{u}(t), \qquad \mathbf{x}(t_0) = \mathbf{x_0} \in \mathbb{R}^n$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector and $\mathbf{u}(t) \in \mathbb{R}^m$ is the control vector.
The optimization problem is to find the control $\mathbf{u}^*$ that minimizes the cost function :

$$J_{LQR} = \frac{1}{2}\int_{t_0}^{t_f} (\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{u}^T\mathbf{R}\mathbf{u})dt \qquad (9)$$

where $\mathbf{Q}$ and $\mathbf{R}$ are the weight matrices.

State Dependent Riccati Equation The feedback optimal solution of the above problem $\mathbf{u}^*$ is given by

$$\mathbf{u}^*(\mathbf{x}) = -\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}^T(\mathbf{x})\mathbf{P}(\mathbf{x})\mathbf{x} \qquad (10)$$

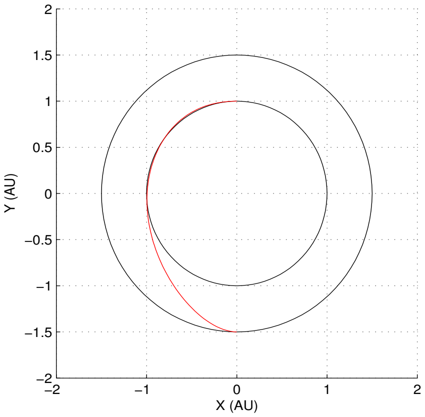


Fig. 1. Trajectory of Earth-Mars transfer in polar coordinates, from [$r_0 = 1, \theta_0 = \pi/2, \dot{r}_0 = 0, \dot{\theta}_0 = 1$] to [$r_f = 1.5, \theta_f = 3\pi/2, \dot{r}_f = 0, \dot{\theta}_f = 0.54433$]

Where $\mathbf{P}(\mathbf{x})$ is obtained by solving the SDRE State Dependent Riccati equation:

$$\dot{\mathbf{P}}(\mathbf{x}) + \mathbf{A}^T(\mathbf{x})\mathbf{P}(\mathbf{x}) + \mathbf{P}(\mathbf{x})\mathbf{A}(\mathbf{x}) + \mathbf{Q}(\mathbf{x}) - \\ \mathbf{P}(\mathbf{x})\mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}^T(\mathbf{x})\mathbf{P}^T(\mathbf{x}) = 0 \qquad (11)$$

We note that the Riccati matrix, $\mathbf{P}(\mathbf{x})$ depends on the choice of $\mathbf{A}(\mathbf{x})$, and since $\mathbf{A}(\mathbf{x})$ is not unique we have multiple optimal solutions.

VII. Optimal Orbit Transfer

Optimal orbit transfer (Numerical examples) In the first example we would like to make an optimal Earth-Mars transfer(i.e. from $(r = 1)$ to $(r = 1.5)$) in time $t_f = 4.469$ (295.8 days). The initial angle is $(\theta_0 = \frac{\pi}{2})$ and the final angle is $(\theta_f = \frac{3\pi}{2})$. $\dot{r}_0 = 0$ and $\dot{r}_f = 0$ for the initial and final orbits. $\dot{\theta}_0 = \sqrt{\frac{1}{r_0^3}} = 1$ and $\dot{\theta}_f = \sqrt{\frac{1}{r_f^3}} = 0.54433105395$. In the second $\theta_f = \frac{5\pi}{2}$ with $t_f = 6.866$(397.6 days).
in both examples the matrices $\mathbf{Q}$ and $\mathbf{R}$ are the identity matrices.

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

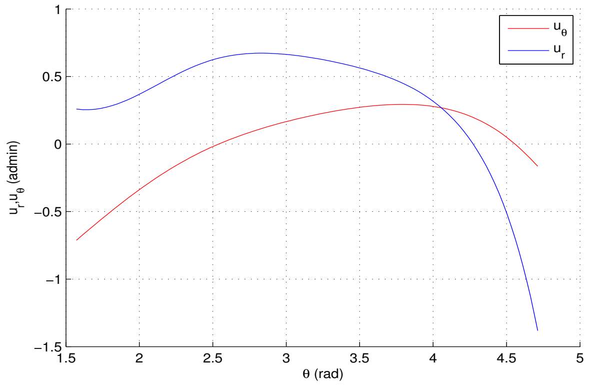


Fig. 2. Controls of the Earth-Mars transfer from [$r_0 = 1, \theta_0 = \pi/2, \dot{r}_0 = 0, \dot{\theta}_0 = 1$] to [$r_f = 1.5, \theta_f = 3\pi/2, \dot{r}_f = 0, \dot{\theta}_f = 0.54433$]

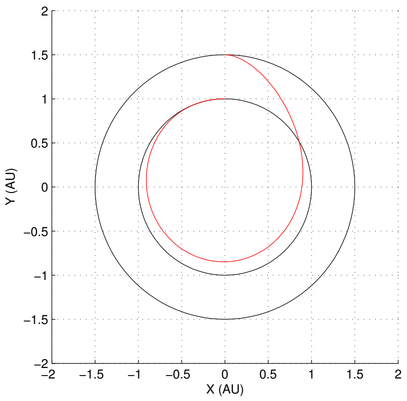


Fig. 3. Trajectory of Earth-Mars transfer in polar coordinates, from [$r_0 = 1, \theta_0 = \pi/2, \dot{r}_0 = 0, \dot{\theta}_0 = 1$] to [$r_f = 1.5, \theta_f = 5\pi/2, \dot{r}_f = 0, \dot{\theta}_f = 0.54433$]

VIII. Conclusion

- The nonlinear feedback optimal control can be solved by factorizing the original nonlinear dyanmics into accessible (weakly controllable) linear dynamics of state dependent factors.
- The factrorized problem has been solved using the SDRE
- The method has been implemented to solve feedback optimal control of Earth-Mars orbit tansfer problem.
- The result is valid for any planet to planet transfer.

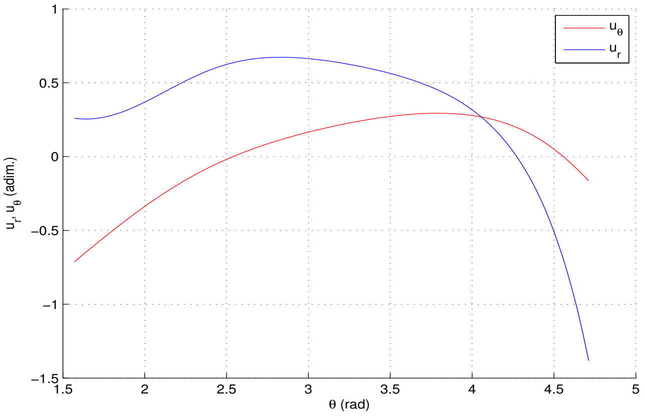


Fig. 4. Controls of the Earth-Mars transfer from [$r_0 = 1, \theta_0 = \pi/2, \dot{r}_0 = 0, \dot{\theta}_0 = 1$] to [$r_f = 1.5, \theta_f = 5\pi/2, \dot{r}_f = 0, \dot{\theta}_f = 0.54433$]

IX. Acknowledgments

This project was supported financially by the Science and Technology Development Fund (STDF), Egypt, Grant No 1834

X. References

1) L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley & Sons, New York, 1962.
2) A. Bryson and Y. Ho, Applied Optimal Control. John Wiley & Sons, New York, 1975.
3) J. Betts, *Practical Methods for Optimal Control using Nonlinear Programming*, SIAM, 2000.
4) P. Enright and B. Conway, *Discrete Approximations to Optimal Trajectories Using Direct Transcription and Nonlinear Programming*, Journal of Guidance, Control, and Dynamics, Vol. 15, pp:994-1002, 1992
5) C. Park and D. Scheeres, *Solution of Optimal Feedback Control Problems with General Boundary Conditions Using Hamiltonian Dynamics and Generating Functions*, Automatica, Vol. 42, pp:869-875, 2006
6) C. Park, V. Guibout, and D. Scheeres, *Solving Optimal Continuous Thrust Rendezvous Problems with Generating Functions*, Journal of Guidance, Control, and Dynamics, Vol. 29, no. 2, pp:321-331 ,2006
7) P.Gurfil *Nonlinear Feedback Control of Low Thrust Orbital Transfer in a Central Gravitational Field* Acta Astronautica, Vol. 60, pp:631-648, 2007
8) F. Topputo, A. Owis, and F. Bernelli-Zazzera *Analytical Solution of Optimal Feedback Control for Radially Accelerated Orbits* Journal of Guidance, Control, and Dynamics, Vol. 31, No. 5, pp:1352-1359, 2008
9) K. D. Hammetty,C. D. Hallz, and D. B. Ridgelyx,*Controllability Issues in Nonlinear State-Dependent Riccati Equation Control*, Journal of Guidance, Control, and Dynamics vol.21 no.5, PP:767-773, 1998