



Cairo University

Institute of Statistical Studies and Research



THE ANNUAL CONFERENCE
ON STATISTICS, COMPUTER SCIENCE
AND OPERATIONS RESEARCH



COMPUTER SCIENCES



26-29, Dec ,2011

Index Computer Sciences

1	The Relationship between Strategy Mapping and the Strategy Formulation, Communication and Implementation : A proposed New Model and an applied e-learning Strategy Map to MSA University Heba M. Adel	1- 22
2	Evaluating Energy Consumption Efficiency of the Zone-based Routing protocol Fahmy. I , Nassef. L , Hefney H.A	23- 32
3	Towards Building Annotated Arabic Web-based Corpus Ezzat .M , ElGhazaly .T	33-41
4	An Enhanced framework for Measuring ERP Implementation Success In the Egyptian Environment Esmat M.Abdelmoniem , Sherif A. Mazen Ehab E.Hassanein	42-56
5	The Role of The pre-development Activities in Information Systems projects"Failure Mohamed A.Hamada , Sherif A.Mazen Ehab E. Hassanein	57-72
6	Arabic Spatial Query Expansion Based on Egypt Geo-Ontology Alaa Hedib , Akram Salah	73-86
7	Using Modularity With Rough Decision Models Ahmed T.Shawky , Hesham A.Hefny, Ashraf H. Abd-Elwahab	87-99
8	A Flexible Data Integration Framework based on Semantic Web Engineering Ibrahim M. El- Henawy , Mahmoud M. AbdEllatif , Tamer E .Amer	100-113
9	Enhancing Text Information Extraction in Videos Using WordNet library (AFS), Sherif F.younis , Fatma A.El- Licy, Akram I. Salah	114-126
10	Cloud Community Framework for Web Services Using Ontology Tamer M. Al-Mashat , Fatma A. El- Licy , Akram I. Salah	127-138
11	Rhetorical Representation for Arabic Text Ahmed Ibrahim , Tarek Elghazaly	139-146

The Relationship between Strategy Mapping and the Strategy Formulation, Communication and Implementation: A proposed New Model and an applied e-learning Strategy Map to MSA University

Heba M. Adel¹

Abstract

Business strategies are difficult to be described, clearly understood, easily communicated, efficiently implemented and effectively evaluated in order to be improved. Thus, accelerating the need for a common visual language used by every one inside and outside the organization, for an effective and efficient strategy communication, implementation and improvement. However, strategy can't be efficiently implemented if it can't be clearly understood, and it can't be clearly understood if it can't be well described. Accordingly, you must first describe your chosen strategy to have a unified clear interpretation of that strategy, then measure the strategy for what needs to be implemented and finally you can improve that strategy by your own measurements. So we should have a systematic continuous cycle of improvement for our strategic management and communication process and in order to achieve this we must first construct a reliable and logical cause and effect framework to describe and improve our chosen strategy.

Based on reviewing the previous literature, a Balanced Score-card (BSC) Strategy Map is needed to maintain such framework which will be introduced throughout this research paper. The strategy map is a visual presentation of an organization's critical success factors and the cause-and-effect relations between them. Strategy maps can be used to depict any organizational strategy through a comprehensive integrated framework of tangible and intangible assets and outcomes. Additionally, Digital technologies are clearly altering the strategic perspectives and operational context of many organizations, especially the educational institutions working in Dynamic markets characterized by oscillating stochastic & uncertain Demand, in an ever-changing fierce globalized competitive Virtual Environments. Thus, accelerating the need for conducting an e-learning system. The main objective of this research is to study how e-learning strategy maps can be used in decreasing the gap between e-learning strategy formulation and e-learning strategy implementation from two opposite directions for the purpose of better e-learning strategy communication and improvement.

This research can be characterized as a qualitative and quantitative research. The empirical part of this research consists of one case applied to an Egyptian private educational institution as a suggested framework for a Continuous e-learning Strategy Communication and Improvement Model, after reviewing many strategy maps designed by different educational institutions in many developed countries such as European, American and Asian countries.

Keywords: e-learning Strategy; Strategy Mapping; Strategy Formulation; Strategy Communication; Strategy Implementation; e-learning Strategy Improvement; Moodle e-Learning Management System; Virtual learning environment; Balanced Score-card; Quality Management of Higher Education.

¹ Lecturer in Modern Sciences and Arts (MSA) University, Faculty of Management Sciences. The author published this research in Cairo University, Faculty of Commerce, Business Administration Department, July 2010, under the supervision of Professor Dr. Ahmed Fahmy Galal.

1. Introduction

Due to the complex nature of business strategy that makes it difficult to be well described, commonly interpreted and easily communicated, most of the strategies fail in the implementation phase. Accordingly, Scholey (2005) pointed out that the Strategy map can help organizations overcome this communication difficulty by providing these organizations with an opportunity to describe and communicate their different strategies in various picture formats. He added that it also helps in achieving better top-down communication, which improves each organizational member's understanding, awareness and interpretation of that chosen strategy and allows for better strategy implementation and improvement.

This research studied how strategy mapping can be used in reducing many of these communication issues through the use of something most people can understand, which is the picture that can tell the organizational strategy in the form of a visual story. By this way, it closes the gap between strategy formulation and implementation through better strategy communication. In addition to that, this research discussed the role of the strategy map as an effective performance management tool that evolved after the development of various performance measures, especially in the educational sector. Moreover, this research investigated how an e-learning strategy map can act as a link between e-learning strategy implementation and formulation (through the e-learning strategy evaluation phase) for the purpose of e-learning strategy improvement, as we'll see in the e-learning strategy communication and improvement model introduced by this research paper.

This suggested model acts as a closed-loop for continuous strategy communication and improvement that bridges the gap between the strategy formulation and implementation from two ways: strategy communication (from one side) and strategy evaluation (from the other side) through a non stop process of improvement. It can be compared to a strategic circulatory system for pumping the lifeblood -which can be viewed in the real life organizations in the form of newly formulated organizational strategies- of any organizational success. This strategic circulatory system is an important organ in any living successful organization. It communicates and transfers the formulated elements of any chosen strategy to each cell in each department at each organizational level preventing many diseases resulting from ineffective implementation of that chosen strategy. It helps stabilizing the organizational temperature to achieve the overall organizational vision through effective implementation of that strategy aligned with the organizational mission statement. Then it carries the feedback after the effective implementation of that strategy to the main senders in order to be clearly evaluated for the purpose of continuous improvement. By this way, the strategy can be pumped upward and downward in each level of the organization to provide the required nutrients (tangible and intangible assets and outcomes) for enhancing the overall organizational performance.

1.1 The Research Problem

The gap between the strategy formulation phase and the strategy implementation phase -represented by ineffective strategy communication and improvement- is a subject that has concerned both strategy academics and practitioners for many years. Different ways and tools have been developed for bridging this gap. This research paper is concerned about one of these tools – the strategy map, especially e-learning strategy map.

However, in the current strategy literature, there is a lack of a detailed model addressing the relationship between the strategy mapping and the strategy formulation, communication, implementation and improvement. Thus, this research studied the opportunity of integrating the strategic management process with the communication process into one closed-loop model, which has been applied through out this research to one Egyptian private educational institution, as a suggested framework for e-learning continuous improvement that can be generalized, later on, by many strategic practitioners and academics in different organizations and various industries.

Additionally, it addresses the effect of the strategy map on helping the educational institutions and universities in measuring and managing their performance, as this is related directly to their chosen strategies and how effectively and efficiently these strategies are communicated, implemented and improved.

1.2 The Importance of the Research

The importance of this research lies in its contribution to the literature of the strategic and performance management from a developing country prospective (Egypt), as the Balanced Scorecard strategy map has been implemented in many developed countries such as European, American and Asian countries for many years but has been ignored in the Arab world especially in Egypt.

Moreover, this research introduced a model that integrates the strategic management process along with its three main phases (formulation, implementation and evaluation) with one of the most popular communication models for the purpose of continuous e-learning strategic improvement.

Additionally, this research will capture the attention on applying the strategy mapping to an Egyptian university, as the BSC strategy map has been used in the business sector but not widely applied to the educational sector especially in Egypt.

The application of strategy mapping to an Egyptian University can help in measuring and managing the performance of this university as it relates directly to its chosen strategy through the strategy evaluation phase. Moreover it can help in implementing its chosen strategy effectively as more attention is always paid to strategy development or formulation than strategy implementation and improvement (e.g., e-learning Strategy).

1.3 The Research objectives

- (A) Using the e-learning Strategy Map as a common visual language for Communication
- To capture the attention of leaders who craft each organizational e-learning strategy to do a better job at communicating that strategy to the "doers" of the organization to help make it real and tangible at their levels.
 - To study the role of the e-learning strategy mapping in linking the strategy formulation phase with the strategy implementation phase through the use of cause-and-effect diagrams in depicting and visualizing each strategy (i.e. e-learning strategy communication).
 - To study the effect of the strategy mapping on the strategy execution phase as many strategies fail in the execution stage as a result of communication

issues. Strategy mapping alleviates many of these communication issues through the use of pictures, something everyone can commonly understand.

(B) Using the e-learning Strategy Map as a common visual language for Improvement

- To study the role of the e-learning strategy mapping in linking the strategy implementation phase once more with the strategy formulation phase but in the opposite direction through the use of interdependent financial and non financial indicators for the purpose of strategy evaluation (i.e. e-learning strategy improvement).
- To show the impact of the strategy maps on closing the huge gap between the vision, mission and different strategies formulated at the top, and between the same strategies but implemented differently at lower levels down in the same organization, through building the bridge between these three initiatives: e-learning strategy implementation, employee empowerment and feedback for continuous improvement especially in the educational sector.
- To find the effect of the strategy map on ensuring that a balanced scorecard is linked to each organization's strategy, so that we can reach full alignment to the overall organizational vision and mission statements.

(C) Integrating the strategic management system of one Egyptian private educational institution with a two-ways communication process model for the purpose of continuously improving its newly formulated strategies (e.g. e-learning strategy and sustainable interdependent cross-cultural strategy).

1.4 The Research Hypotheses

In order to achieve these research objectives, the following hypotheses were developed:

H 1: The more the understanding of e-learning strategy formulation elements, the better the e-learning strategy can be described through strategy mapping as a set of logical cause and effect relationships between different strategy activities (a set of drivers for desired outcomes).

H 2: The better the e-learning strategy can be described as a set of logical cause and effect relationships between different strategy activities, the more unified understanding and clear interpretation of that e-learning strategy among all levels.

H 3: The less the differences in individuals' interpretation of an organization's e-learning strategy, the easier and more effective e-learning strategy implementation will be.

H 4: This hypothesis is divided into two sub-hypotheses as follows:

H 4/1: The better the financial and non financial performance indicators can be linked into measurable cause and effect framework through e-learning strategy mapping, the easier the e-learning strategy can be evaluated.

H 4/2: The easier the e-learning strategy can be evaluated from different perspectives, the more the chosen e-learning strategy can be improved.

These research hypotheses were developed based on what Norton (1999) have discussed about a set of premises upon which the Balanced Scorecard (BSC) design process is built, and one of these premises stated that strategy can be described as a set of cause-and-effect relationships between different strategy activities that are the drivers of desired outcomes.

1.5 The Research Models

1.5.1. Research Graphical Model

Closed-loop Model for Continuous
e-learning Strategy Communication and Improvement

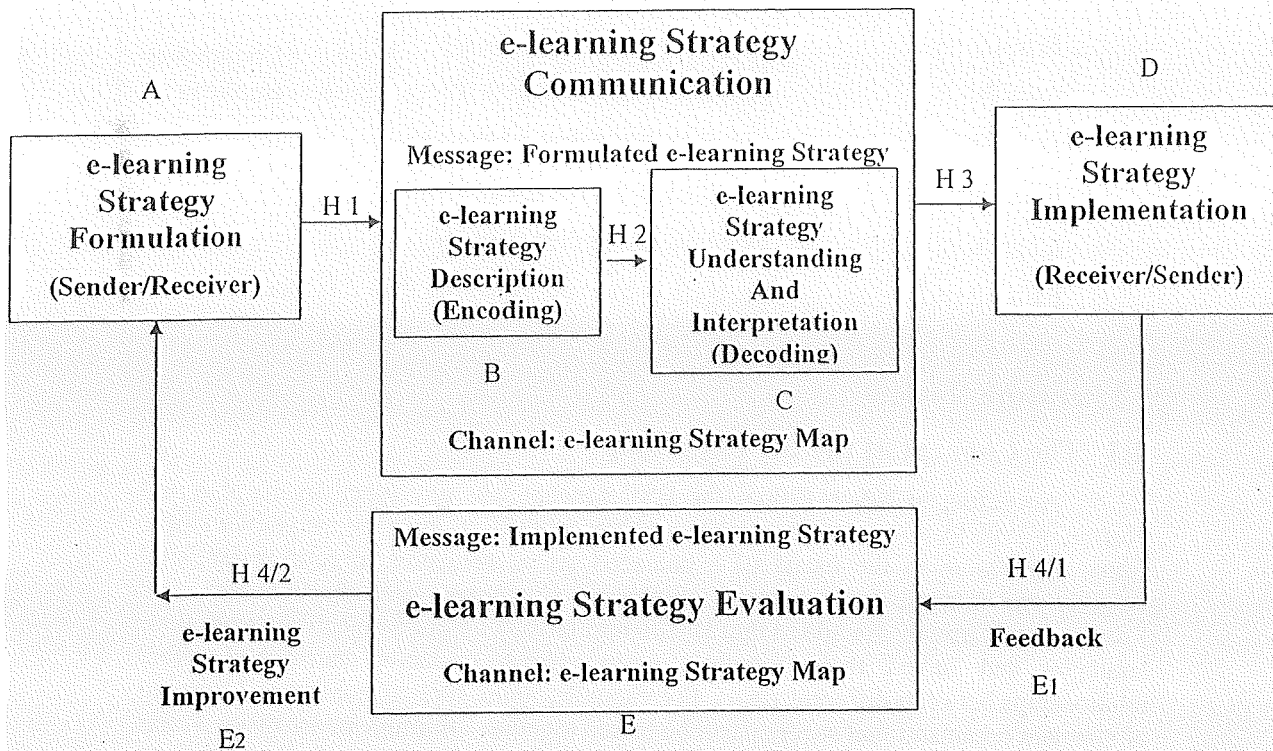


Figure (1): This Research Graphical Model

1.5.2. Research Verbal Model

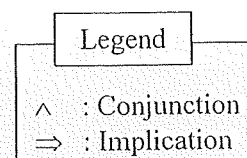
Based on Schramm's (1955) interactive model of communication (cited in belch & belch, 2007), a continuous e-learning strategy communication process needs basic elements to be effective:

- Source/Sender: Everyone responsible for the e-learning strategy formulation.
- Receiver: Everyone responsible for implementing that e-learning formulated strategy at each organizational level, department, branch and unit.
- Message: The chosen formulated e-learning strategy through the e-learning strategy communication and the implemented e-learning strategy through the e-learning strategy evaluation.
- Channel: The e-learning strategy map as a visual communication and performance management tool.
- Encoding: Clear e-learning strategy description.
- Decoding: Unified e-learning strategy understanding and interpretation.

- g. Feedback: Through the e-learning strategy evaluation which is later on transformed into strategy improvement. This could be achieved when the receiver in the e-learning strategy implementation act as the sender for communicating the feedback of the evaluated e-learning strategy back to the source in the e-learning strategy formulation.

1.5.3. Research Mathematical Logical Model

$A \Rightarrow B$	i.e. if A is true then B is true
$B \Rightarrow C$	i.e. if B is true then C is true
$C \Rightarrow D$	i.e. if C is true then D is true
$D \Rightarrow E1$	i.e. if D is true then E1 is true
$E1 \Rightarrow E2$	i.e. if E1 is true then E2 is true
$E2 \Rightarrow A$	i.e. if E2 is true then A is true
$E1 \wedge E2 \Rightarrow E$	i.e. if E1 is true and E2 is true then E is true



2. Literature Review

2.1 The Concept of the Strategy Map

According to Niven (2005) the BSC concept was developed by Kaplan and Norton in 1990 to solve a measurement problem by shifting from the usage of a single financial measure to the usage of a balanced set of measures in four interrelated tangible and intangible perspectives: Financial, Customer, Internal Processes, and Learning and Growth. He added that BSC teams have developed objectives from their strategies and linked them together, using arrows to draw cause and effect relationships between different items. For example, items such as quality and training of a certain organizational strategy can be linked together - through an arrow on a map or chart- indicating that providing competency based training to its employees in its never ending continuous process of improvement will allow them to produce quality featured products or services with fewer defects. Finally, the strategy map was born, and as Kaplan and Norton pointed out, it was created to be as important innovation as the original BSC itself (cited in Niven, 2005). According to Niven (2005) the strategy map is a picture that tells a thousand words.

Norton (1999) defined the Strategy Map as a pictorial cause and effect logical reliable and consistent architecture. He added that it represents a common model of the organization's strategy to help translate its overall quality vision into action and links the four-tier BSC framework with the strategic themes. By clearly defining near-term objectives and activities, it builds and sustains long-term customer partnerships and shareholder value.

In a similar vein, De Waal (2003) viewed the Strategy map as graphical comprehensive cause and effect generic framework that describes a strategy by specifying the relationships between different stakeholders, internal and external customers, business processes and core competencies. In addition to that, Kaplan and Norton (2000) argued that this strategy map's cause-and-effect links reveal how specific improvements create desired outcomes. For example, a strategy map of an

educational institution can show how competency based training to its teaching staff will improve their knowledge, skills and abilities (learning and growth perspective), enhance their learning, teaching and assessment capabilities (internal process perspective), and exceed the expected level of quality perceived by different stakeholders (customer perspective), which will finally expand the organization's revenues and increase its assets utilization (financial perspective). From a larger perspective, strategy maps depict how an organization will transform its intangible assets such as positive cultural values, effective leadership and communication skills into tangible outcomes. Moreover, it acts as an effective performance management tool as we will see through out the research paper.

2.2 Using the Strategy Map as a common visual language for communication

As for how the strategy map can visually tell the story of an organization's strategy, Amarangtunga et al. (2001) indicated that the strategy map as a graphical representation of the causal integrated model of an organization's strategy tells the story of that strategy (cited in Othman, 2006). In addition to that, Cummings and Wilson (2003) argued how the strategy map can also empower each one responsible for each activity at each level in the organization to do his own mapping, which creates a motivation for bringing other opinions to the fore. It may not get people down the mountain in a straight line but it gets things moving and, therefore, other things come into view (cited in Thomas, 2007).

In a similar vein, Kaplan and Norton (2004) discussed the usage of the strategy map as a common language for strategy communication through pointing out how the strategy map can express graphically the cause-effect links between different strategy components. Moreover, it acts as a platform for high-level managers to deliver the message of organization's direction using unified visual language for creating business value (cited in Huang and Lee, 2006).

Kaplan and Norton (2000) came to another realization that the key to an effective and efficient strategy implementation in any organization is through allowing its employees to use strategy mapping so that they can clearly view how their jobs are adding value toward achieving the overall organizational objectives. These strategy maps can enable them to work in a state of interdependence and synergy toward the organization's shared quality vision, mission and core values.

From their experience, Kaplan and Norton (2000) have developed a standard template that organizations can use in developing their own strategy maps in order to close the huge gap between the formulated strategy at the top and the implemented ones down the organization. That template contains four different parts -financial, customer, internal process, and learning and growth-that correspond to the four perspectives of the balanced scorecard.

2.3 Using the Strategy Map as a common visual language for improvement

This research investigated how the strategy map can act as a link between strategy implementation and formulation for the purpose of strategy improvement as we have seen in the strategy communication process model. In addition to that, this research discussed the role of the strategy map as an effective performance

management tool that evolved after the development of various performance measures.

Based on reviewing the literature on the progress of performance measurement development, three lines of research or phases can be distinguished. The first one focused on the development of the single measure (cost and management accounting, 1850-1925), the second one focused on the development of the multiple measures (multi-dimensional performance measurement frameworks, 1974-1992), while the third one focused on the development of the BSC and Strategy maps (business models or cause and effect diagrams, 1992-2000) (Wilcox and Bourne, 2003).

In the Industrial age, organizations created value by transforming raw materials into finished products. The economy was based mainly on tangible assets - equipment, factories, land, and inventory - and any organization needed to describe and measure, its business strategy had no options except using financial tools such as balance sheets, income statements, and general ledgers as the traditional performance measurement grew out of accounting (Kaplan and Norton, 2000).

However, by the late 1980s it was evident that the use of financial performance measures alone was inadequate for the management and improvement of modern organizations that led to the development of multi-dimensional performance measurement frameworks through which non financial performance indicators lead financial indicators (Wilcox and Bourne, 2003).

Later on, in the Information age and after the development of the BSC, strategy maps were proven to translate the concept of "leading and lagging indicators" into a testable cause and effect generic architecture where each tangible and intangible performance measure is linked to another, so that organizations can build a sustainable competitive advantage in a globalized changing business environment through creating and deploying intangible assets-for example, customer supplier partnerships; employee knowledge, skills and abilities; information technologies; and a new culture value system created to encourage innovation, creativity, problem solving, and total continuous process of improvements (Kaplan and Norton, 2000, Wilcox and Bourne, 2003).

2.4 Applying Strategy Mapping to the Educational sector

Kaplan and Norton (2000) have introduced strategy maps into various industries, including retail, insurance, banking, telecommunications, health care, chemicals, e-commerce, and energy as well as some non profit organizations and government units, but these strategy maps were not widely applied to the educational sector.

However, some studies were made concerning the usage of strategy maps in the educational sector; for example, Chen, yang and shiau (2006) discussed the need for using these strategy maps in Taiwanese higher education. They pointed out that the Taiwanese educational sector has been changed because of the rapid expansion in university education. In addition to that, they came to a realization that this change can be seen in the unbalanced supply and demand in university education as well as the reduction in the educational quality. They added that as a result of the fierce international competition in educational services, many countries have invested more in university education to support their quality improvement efforts nationally and build a sustainable competitive advantage internationally. Based on that realization,

they investigated the usage of a BSC Strategy map in improving the quality of the educational programs as well as enhancing the overall operational performance in university education.

In a similar vein, Smeby (2003) pointed out that universities and other institutions of higher learning are facing the challenge of being centers of excellence for teaching, learning and research. On one hand universities are increasingly being asked to teach an increasing number of students in increasing numbers of specializations, and on the other hand they are being forced to pay more attention to quality of teaching, educational programs and other internal educational processes as well as the quality and performance of its teaching staff (cited in Umashankar and Dutta, 2007). According to Umashankar and Dutta (2007) the BSC Strategy map can provide Indian universities and other institutions of higher education with an opportunity to build a cascade of measures to translate the knowledge management and other intangible assets into a comprehensive integrated clear coherent and communicable framework for positioning their students and programs positively in the minds of external stakeholders and international audience.

Based on reviewing the previous literature, some studies focused the attention on the BSC Strategy Map, and tried to discuss how it may be used as one of the effective performance and strategic management tools in higher educational institutions and universities. It can set educational measures and objectives for the purpose of improving the effectiveness, efficiency and competitiveness of the educational process in an ever changing globalized environment. Moreover, it enhances the efficiency of delivering the educational quality message to every one inside and outside these educational institutions.

This research introduced some of these models from different countries in public and private educational institutions which are using the strategy mapping in improving their strategic performance management system.

- 2.4.1 An American perspective (Papenhausen and Einstein, 2006)
- 2.4.2 A Taiwanese perspective (Chen et al., 2006)
- 2.4.3 An Indian perspective (Umashankar and Dutta, 2007)
- 2.4.4 A British perspective (Cullen et al., 2003)

2.4.1 An American Business College

Papenhausen and Einstein (2006) stressed the importance of the BSC Strategy map for both the strategy communication and evaluation by discussing the role of the strategy map in transforming the intangible assets into tangible customer and financial outcomes. They argued how the strategy map can root each strategy measure in a chain of cause and effect logic that relates the desired strategic outcomes with the drivers that will lead to these outcomes. Finally they concluded that the successful implementation of any strategy needs active collaboration and synergy by every one in the organization. They added that each college member needs to clearly understand and interpret that chosen strategy through successful strategy communication which is one of the important factors in achieving these initiatives. Besides a successful strategy evaluation through using the BSC Strategy map can provide each college member with a feedback about his own performance for the purpose of improving the overall college performance.

2.4.2 A Taiwanese University

In a globalized environment that requires increasing responsibility and value adding from different business schools, according to Chen et al. (2006) the BSC strategy map framework not only describes, clearly defines and links different strategic targets but also includes performance measurable indicators of different perspectives for effectively examining and improving the chosen strategy as well as the school's growth.

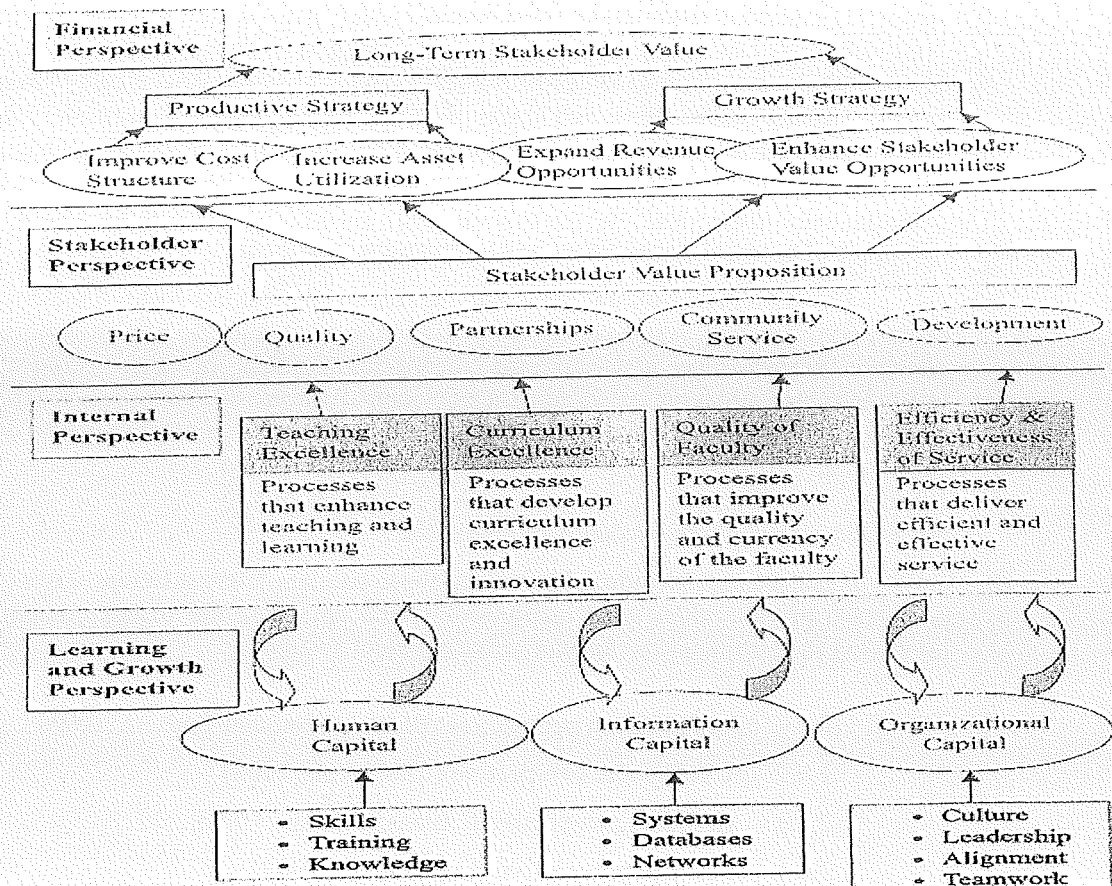


Figure (2): The College BSC Strategy Map based on Kaplan and Norton (2004)

Source: Papenhausen and Einstein, 2006

2.4.3 An Indian University

Pandley (2005) indicated that a well constructed BSC that is easily understood, simply interpreted and systematically managed can be used as an approach for performance management and improvement as well as a mechanism for strategy communication to all levels of management (cited in Umashankar and Dutta, 2007).

Based on the BSC Strategy Map model that is proposed through out this Indian study for different institutions of higher Indian education, we can conclude that

a comprehensive graphical cause and effect strategy map can act as a visual common language for every one inside and outside the organization to clearly understand the organization's strategy and serve as a communicable point of reference for all departments, units and branches to commonly interpret the organization's strategy and at the same time play as a powerful performance management tool that measures and improves every one's performance which is reflected in enhancing the overall organizational performance.

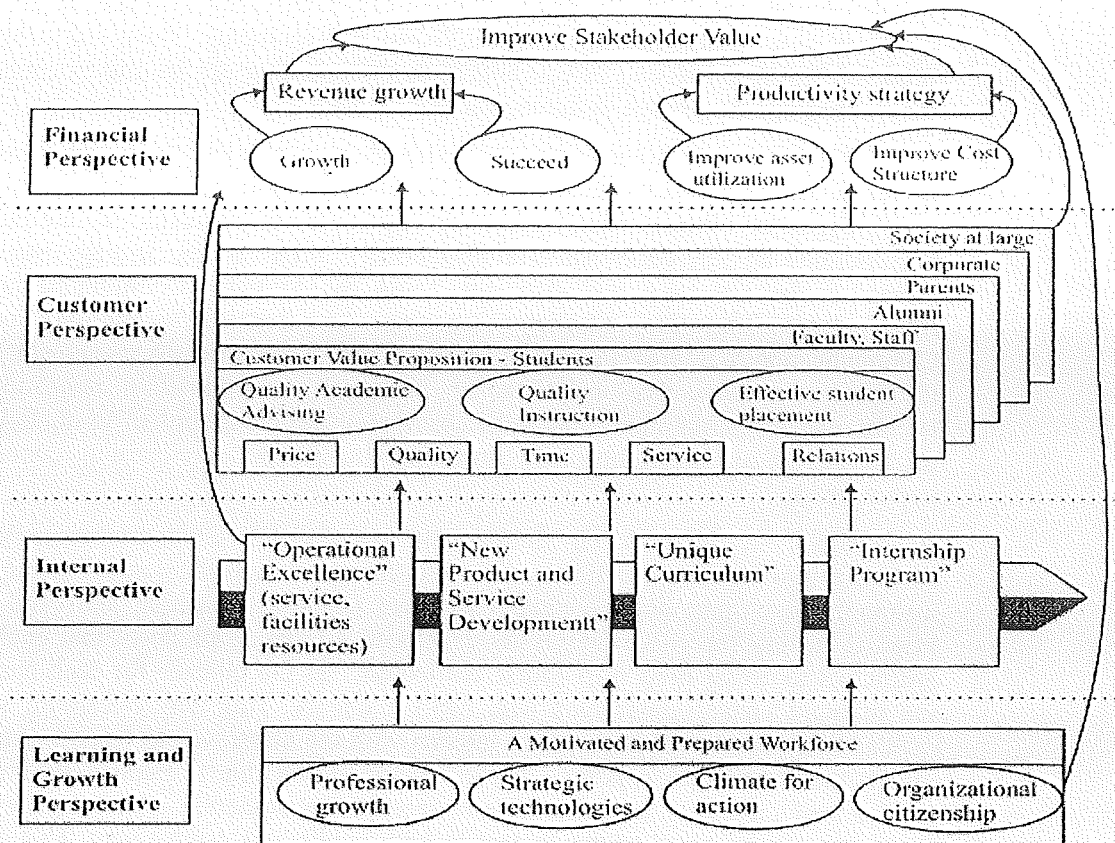


Figure (3): The BSC Strategy Map based on Kaplan and Norton (2001)

Source: Umashankar and Dutta, 2007

2.4.4 A British Business and Management Faculty

A balanced scorecard strategy map for a faculty of management and business at a mid-ranking UK university is constructed. By mapping different strategic actions, we can achieve clarity to the cause and effect issues across the four BSC perspectives and understand the nature of the faculty objectives, through developing a group of interrelated financial and non financial factors as well as tangible and intangible measures that are required to attain the overall organizational mission (Cullen et al., 2003). Moreover, Cullen et al. (2003) recommended further studies to be made regarding the potential use of the BSC and its associated Strategy Map as a mean of moving from monitoring to the management of quality in higher education.

3. A Suggested Strategy Map for an e-learning Strategy Communication and Continuous Improvement applied to an Egyptian Private Educational Institution (MSA University) as a tool for learning and growth

Fry (2001) defined e-learning as the Delivery of Education and Training (KSA) via Networked Interactivity that relies on ICT & CMC, thus, overcoming Time, Space and Physical Limitations in a Virtual Environment. Fry (2001) pointed out that Digital technologies are clearly altering the strategic perspectives and operational context of many organizations, especially those working in Dynamic markets characterized by oscillating stochastic and uncertain Demand, in an ever-changing fierce globalized competitive Virtual Environments. Thus, enhancing the need for having e-learning in educational institutions that are aiming to cope with these changing competitive technological environment.

Moreover, educational institutions should make sure that the four core aspects of Technology -Hardware (Techno-ware), Software (Info-ware), Peopleware (Human-ware), Orgaware (Organization-ware) - are taken into consideration, for the purpose of sustaining an Interdependent Multidisciplinary View. Since, Strategy Maps are proven to be Effective and Efficient Tools for Strategy Communication and Improvement. Therefore, Strategy Maps can be used for e-learning Strategy Communication & Improvement. The empirical part of this research was conducted in the form of a qualitative analysis to an integrated comprehensive closed-loop e-learning strategy communication and improvement process, applied to an Egyptian private educational institution as a suggested framework for e-learning continuous improvement.

MSA e-learning Strategy: Introducing new academic under graduate programs as well as professional certificates delivered through new technological methods (e-learning) for current and potential customers (students, teaching staff and other external customers). MSA University developed e-learning platform that enable its instructors to create their own course curriculum, set the quizzes for such materials and to build virtual classrooms with MSA Students. This new e-learning system enables MSA Student(s) and staff to interact with the classroom as to post questions and get the instructors' response. The instructor will be able to evaluate the students as in a real class, set assignments and receive the students' response, share the comments and guidance to the student in both public and private way. This system emphasizes on the instructor ability to develop and manage the system with the minimal IT skills and the minimal IT professional interaction, in order to maximize the benefits from it, through using the Moodle in a Virtual Learning Environment.

MSA e-learning strategy is delivered through the Moodle. According to Brandl (2005), Riordan and Marcais, Moodle is a Course Management System (CMS), also known as a Learning Management System (LMS) or a Virtual Learning Environment (VLE). Moodle provides a range of functions designed to support the delivery of online or blended learning. It is a world-wide, ongoing development project designed to support a social framework of education. Moodle creates an environment that allows for collaborative interaction among students in addition to virtual classroom instruction. MSA University site first shows a Course listing which shows descriptions for every course on the server. Courses are categorized and can be searched easily. The site can support thousands of courses. Each MSA University teaching staff member has full control over all settings for each course. Each Teacher can choose each course format settings such as by week or by topic. Instructors and

tutors are provided flexible array of course activities: Forums, Quizzes, Glossaries, Resources, Surveys, Assignments, Chats, and Workshops.

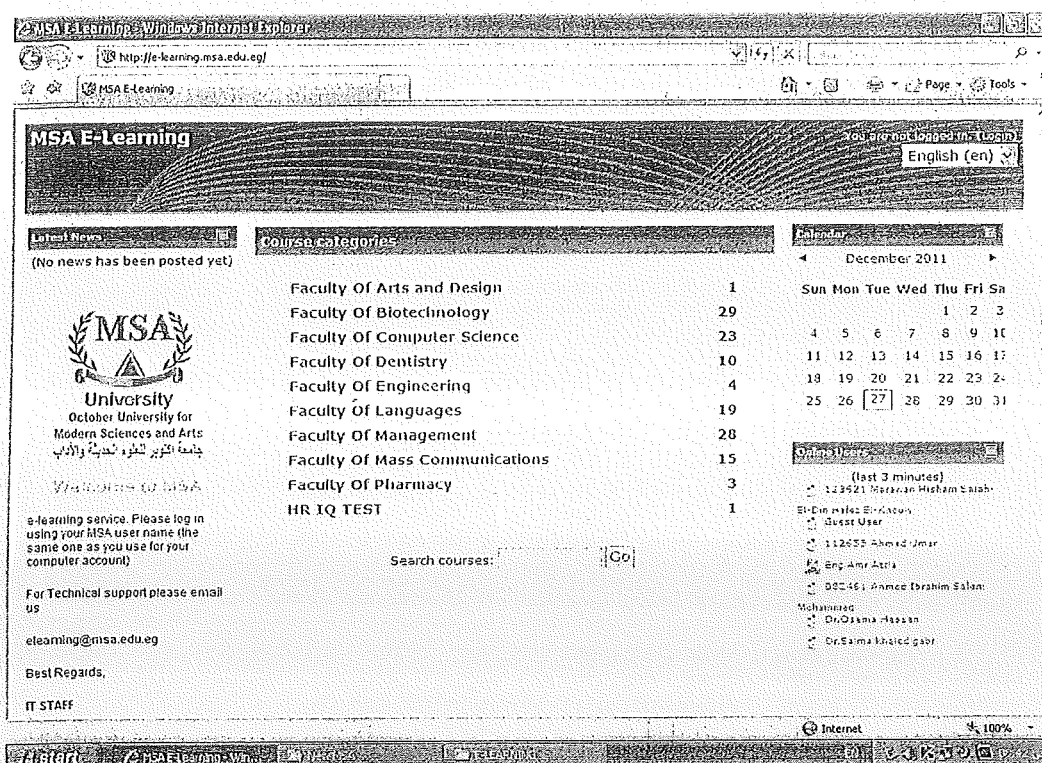


Figure (4): MSA University e-learning, home page

Source: e-learning.MSA.edu.eg

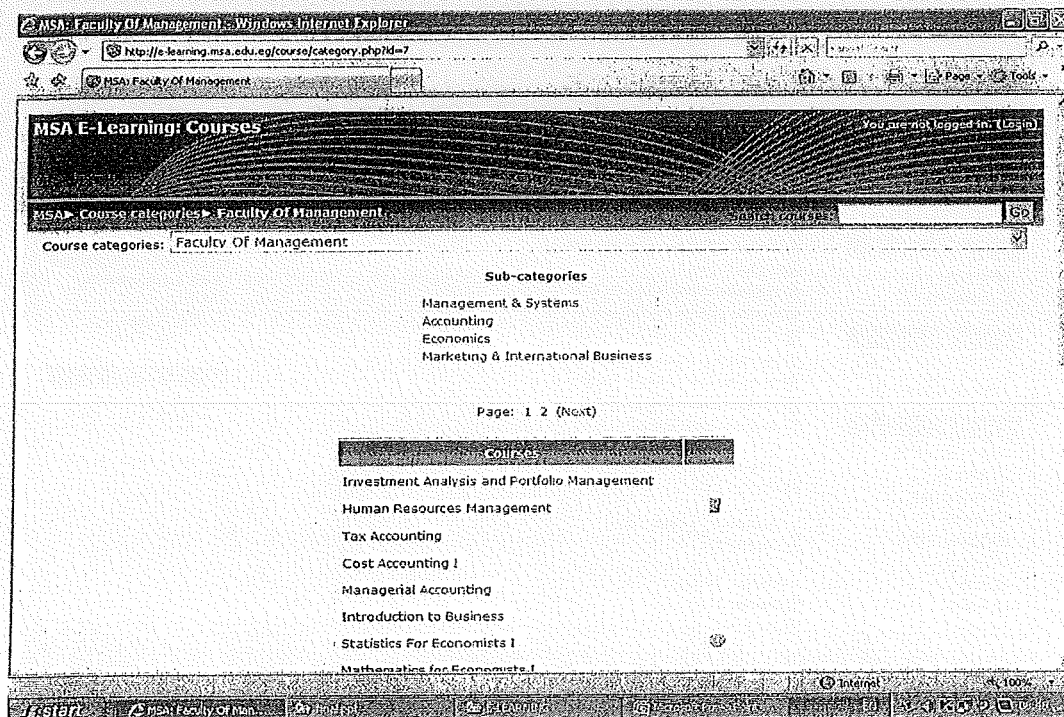


Figure (5): MSA University e-learning, Faculty of Management Sciences Courses

Source: e-learning.MSA.edu.eg

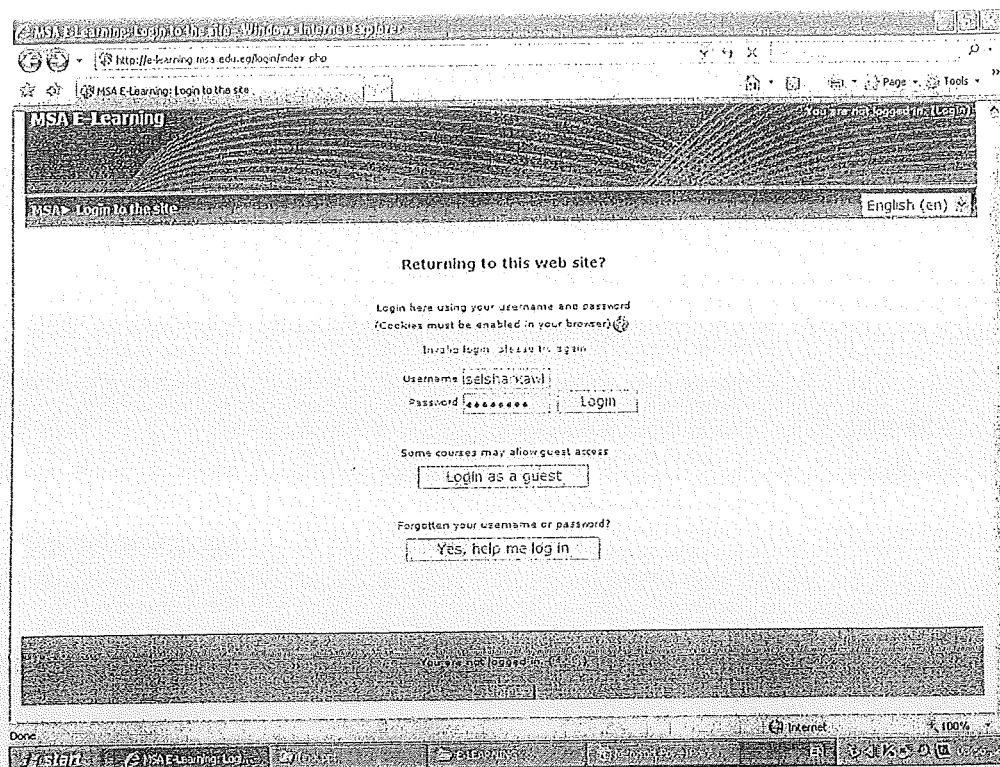


Figure (6): Login to MSA University web site

Source: e-learning.MSA.edu.eg

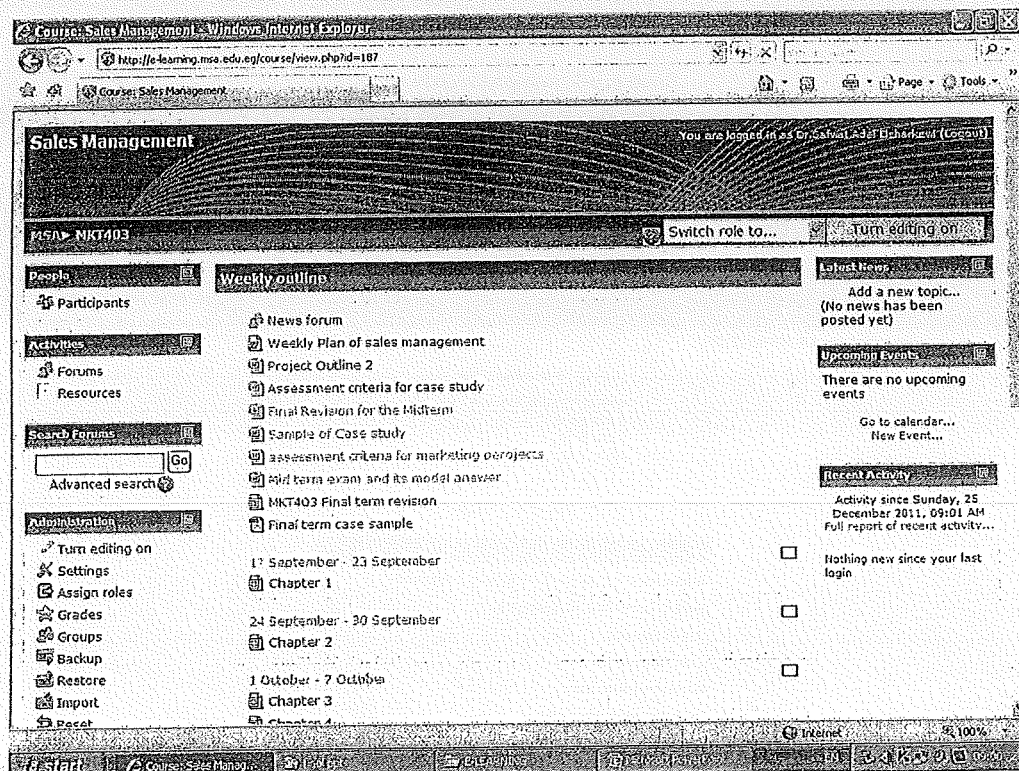
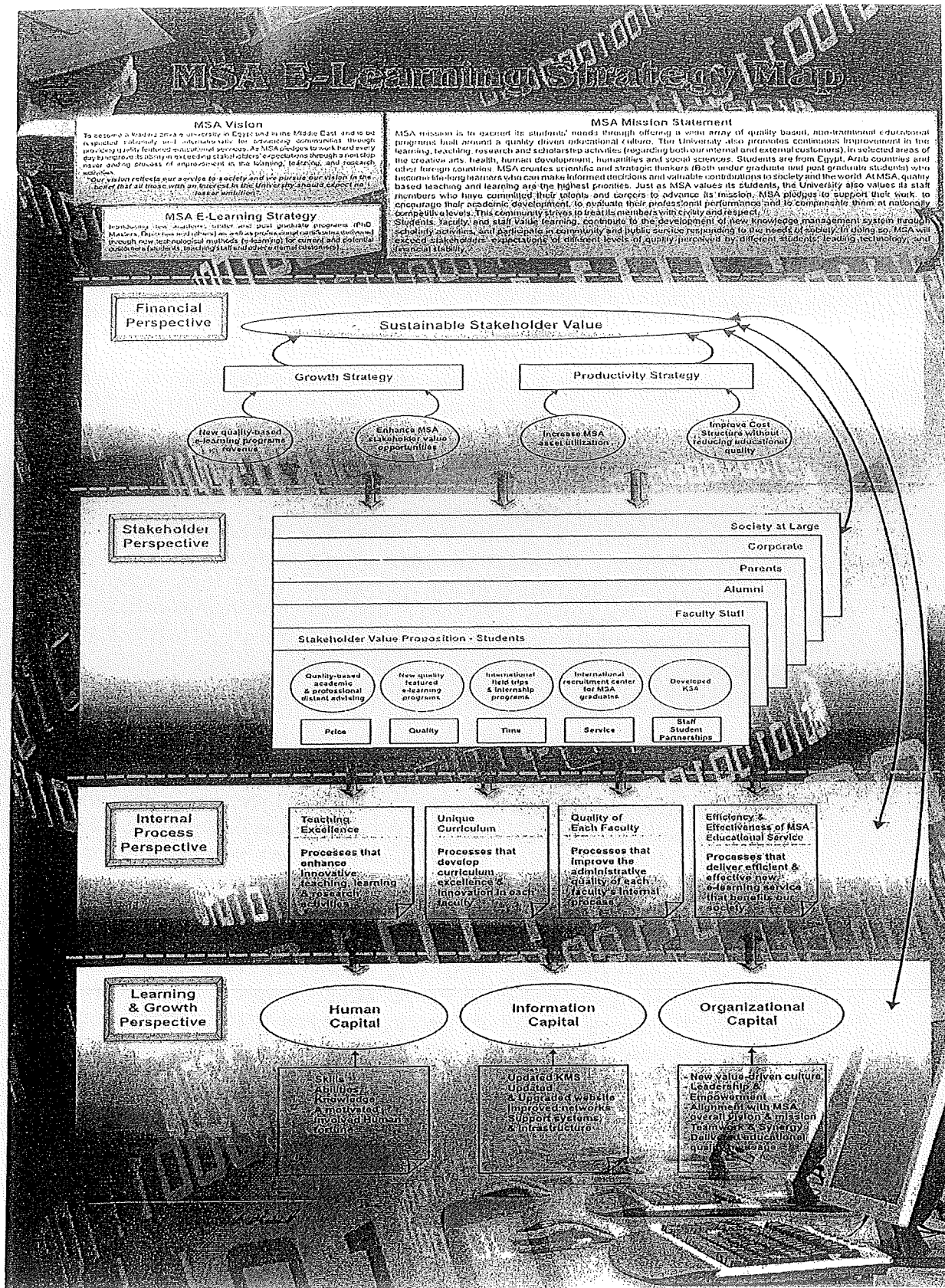


Figure (7): MSA University Sales Management Course materials

Source: e-learning.MSA.edu.eg

Figure (8): MSA e-learning Strategy Map



4. Research Methodology

In order to test the research specific hypotheses and examine the validity of various relationships among different variables included in these research hypotheses, the research design was chosen to be a Conclusive Descriptive research.

This conducted research can be characterized as a qualitative and quantitative research. The qualitative part of this research was applied to MSA University (Egyptian private university), through carrying out a comprehensive strategic audit, to gain a qualitative insight and understanding of the primary data collected through: (a) Depth Interviews conducted with some individuals of MSA leaders, teaching and administrative staff, and (b) Focus Groups held with MSA internal (MSA teaching and administrative staff) and external customers (MSA students and alumni). While, the quantitative part of this research was applied to the same university, to quantify the primary data collected, through personal and electronic (via e-mails) surveys, then a quantitative statistical analysis is used to interpret and analyze this quantitative data.

A single cross sectional design was chosen as the questionnaires will be distributed and collected once, after illustrating the required strategy maps for MSA university by the researcher (e.g. e-learning strategy map and sustainable interdependent cross-cultural exchange strategy map) and conducting strategy mapping workshops for MSA teaching and administrative staff as well as MSA students, for the purpose of giving a common understanding of the importance of strategy mapping to MSA, explaining the integrated closed-loop model of this research and discussing a conducted strategic audit for analyzing MSA strategic position in the Egyptian educational sector. Then, evaluating the changes that took place in the level of MSA strategy understanding and interpretation of its staff, before the strategy maps' workshops were conducted, and their improved performance after these sessions.

4.1. Research Sampling

According to Eisenhardt (1989), case study is a research approach that concentrates on understanding and interpreting the dynamics presented within single settings. A Case study is carried out in order to test and examine in-depth certain topic in a selected case (cited in Antola *et al.*, 2006). These features were one of the main reasons (in addition to the previously stated reasons) for choosing the case research approach to be applied in this study to MSA University.

The Target Population (Universe "N") in this case research includes all MSA teaching and administrative staff working in eight different faculties at MSA University (i.e. 1,106 individuals); the Extent: MSA University; Time: period of survey (2010). The Sample "n": All the teaching and administrative staff working in the Faculty of Management Sciences; the Sampling Unit: Professors, Professors Assistants, Lecturers, Lecturers Assistants, Teaching Assistants and all the administrative staff working in the Faculty of Management Sciences; the Sample Size: 75 individuals. The Sampling Technique: Non probability judgmental Sample (Faculty of Management Sciences) was chosen, based on the researcher judgment and opinion that the individuals working in the Faculty of Management Sciences can better understand the continuous process of strategy communication and improvement presented in this research, based on their educational background and experience.

4.2. Data Collection

In order to test the research hypotheses, two types of data were required:

A. Primary Data collected from:

1. Questionnaires (Surveys): As quantitative descriptive primary data were needed, questionnaires were distributed and collected once, after conducting the strategy mapping process and introducing MSA strategy maps depicted by the researcher, through various strategy maps workshops for the Faculty of Management Sciences teaching and administrative staff.
2. Depth Interviews: As qualitative primary data were required, depth interviews were held with MSA University leaders and decision makers (e.g. Faculty of Management Sciences leaders, MSA library manager, MIS team members).
3. Focus Groups: As qualitative primary data were needed, focus groups were conducted with MSA internal customers (MSA teaching and administrative staff) as well as MSA external customers (MSA students and alumni, especially those students in the Faculty of Management Sciences and Biotechnology).
4. All the strategic matrices and the performance management models that were made by the researcher for this research (e.g. Strategic Groups, Issues Priority Matrix, EFAS, IFAS, Industry Success Matrix, BCG Matrix, TWOS Matrix and other strategic matrices).

B. Secondary Data collected from:

1. Internal data available within the university's Databases, either ready to be used or require further processing (e.g. MSA knowledge Management System).
2. External data: strategic and performance management published materials from the internet websites.

4.3. Measurement and Scaling

The questionnaire developed for this research included 101 questions that covered all the data required for testing the research hypotheses. The Interval Likert scale was used in this research questionnaire as a non comparative itemized rating scaling technique, this measurement scale consists of 5 response categories ranging from "1= strongly disagree" to "5= strongly agree" which requires MSA respondents to indicate a degree of agreement or disagreement of each of the series of statements related to this research questions.

4.4. Data Analysis

Statistical Package for Social Sciences (SPSS) is one of the leading computer packages for analyzing survey data and is the principle package for analyzing data from surveys conducted in the social sciences. Therefore, SPSS was used (Pearson's correlation) to test the Correlation between the variables in each hypothesis, the Reliability and Factor Analysis of these research variables. Quantitative Data analysis was made using the SPSS to analyze the primary data collected from the research questionnaires. While, the Qualitative Data analysis was conducted using different strategic matrices and performance management models that were made by the researcher for this study (e.g. Strategic Groups, Issues Priority Matrix, TWOS Matrix and other strategic matrices).

5. Summary of Findings and Conclusions

The quantitative part of this research was applied to MSA University, to quantify the primary data collected, through personal and electronic (via e-mails) surveys, then a quantitative statistical analysis is used to interpret and analyze this quantitative data. This Quantitative Data analysis was made using the SPSS to analyze the primary data collected from the research questionnaires.

Table (1): Pearson's correlation coefficient

		Correlations					
		A	B	C	D	E1	E2
A	Pearson Correlation	1	.700**	.881**	.866**	.832**	.865**
	Sig. (1-tailed)		.000	.000	.000	.000	.000
B	Pearson Correlation		1	.662**	.804**	.682**	.681**
	Sig. (1-tailed)			.000	.000	.000	.000
C	Pearson Correlation			1	.886**	.868**	.864**
	Sig. (1-tailed)				.000	.000	.000
D	Pearson Correlation				1	.896**	.847**
	Sig. (1-tailed)					.000	.000
E1	Pearson Correlation					1	.847**
	Sig. (1-tailed)						.000
E2	Pearson Correlation						1
	Sig. (1-tailed)						

** Correlation is significant at the 0.01 level (1-tailed).

From the above table, at 0.05 level of significance we reject H₀. On a scale from -1 to 1, the direction and the strength of the relation were measured; the relation is assumed to be weak if the correlation coefficient < 0.5 and strong if more than that. The sign gives the direction of the relation, if positive, the relation is direct and if negative the relation is indirect. Therefore, there is a strong direct relationship between: A & B; B & C; C & D; D & E1; E1 & E2; and E2 & A.

Thus, the following research hypotheses were proven to be accepted (i.e., after rejecting the null hypotheses), it was proven that:

1. The more the understanding of e-learning strategy formulation elements, the better the e-learning strategy can be described through e-learning strategy mapping as a set of logical cause and effect relationships between different strategy activities (a set of drivers for desired outcomes).
2. The better the e-learning strategy can be described as a set of logical cause and effect relationships between different strategy activities, the more unified understanding and clear interpretation of that e-learning strategy among all levels.

3. The less the differences in individuals' interpretation of an organization's e-learning strategy, the easier and more effective e-learning strategy implementation will be.
4. The better the financial and non financial performance indicators can be linked into measurable cause and effect framework through e-learning strategy mapping, the easier the e-learning strategy can be evaluated.
5. The easier the e-learning strategy can be evaluated from different perspectives, the more the chosen e-learning strategy can be improved.

6. Recommendations

This research recommends that the application of strategy mapping to MSA University, should be continued and improved, as it can help in measuring and managing the performance of this university as it relates directly to its chosen strategy through the strategy communication and evaluation phases. Moreover it can help in implementing its chosen strategy effectively as more attention is always paid to MSA strategy development or formulation than strategy implementation and improvement.

Thus, the researcher recommends MSA leaders and managers to:

(A) Use MSA e-learning Strategy Maps as common visual language for Communication, through:

1. Doing a better job at communicating each strategy to the "doers" of MSA to help make it real and tangible at their levels.
2. Linking each strategy formulation phase with MSA strategy implementation phase through the use of cause-and-effect diagrams in depicting and visualizing each strategy (i.e. e-learning strategy communication).

(B) Use MSA e-learning Strategy Maps as common visual language for Improvement, through:

1. Linking MSA strategy implementation phase once more with its strategy formulation phase but in the opposite direction through the use of interdependent financial and non financial indicators for the purpose of strategy evaluation (i.e. e-learning strategy improvement).
2. closing the huge gap between MSA vision, mission and different strategies formulated at the top, and between the same strategies but implemented differently at lower levels down in MSA University, through building the bridge between these three initiatives: e-learning strategy implementation, employee empowerment and feedback for continuous improvement.
3. Ensuring that MSA balanced scorecard is linked to each MSA e-learning strategy, so that we can reach full alignment to the overall MSA vision and mission statements.

(C) Use this research model for communicating and improving each MSA formulated strategy, through integrating MSA strategic management system with its strategy communication process, for the purpose of continuously improving its newly implemented strategies (e.g. e-learning strategy and sustainable interdependent cross-cultural strategy).

7. Suggestions for further research

The researcher recommends further studies and researches to be conducted, in order to generalize and apply this research model to other private and public educational institutions, in addition to other business and industrial sectors.

Specifically, further studies can be developed, so that this research model can be effectively and efficiently assessed when applied to:

1. Different Egyptian sectors, especially the educational sector in Egypt and the Ministry of Education.
2. Other public and private universities, not only in Egypt but also in different countries in the Arab world.
3. Other Business and Industrial sectors, not only in Egypt but also in the Middle East and Arab countries.

References

- Antola, Juha., Lönnqvist, Antti., & Uusi-Rauva, Erkki. (2006). Designing a Strategy Map in order to facilitate strategy implementation. Tampere University of Technology press.
- Belch, G. E., & Belch M. A. (2007). *Advertising and promotion: an integrated marketing communications perspective*, 7th Ed., - Boston: Irwin/McGraw-Hill.
- Brandl, K. (2005). Are You Ready To "MOODLE"?, *Language Learning & Technology*, 9, 2, 16-23
- Chen, S., Yang, C., & Shiàu, J. (2006). The application of balanced scorecard in the performance evaluation of higher education. *The TQM Magazine*, 18, 2, 190-205.
- Cullen, J., Joyce, J., Hassall, T., & Broadbent M. (2003). Quality in higher education from monitoring to management. *Quality Assurance in Education*, 11, 1, 5-14.
- De Waal, A. A. (2003). The Future of the Balanced Scorecard: An interview with Professor Dr Robert S. Kaplan. *Measuring Business Excellence*, 7, 1, 30-35.
- Fry, K. (2001). E-learning markets and providers: some issues and prospects. *Education + Training*, Volume 43, Issue 4/5, Emerald, 233-239.
- Huang, S., Lee, C., & Kao A. (2006). Balancing performance measures for information security management: A balanced scorecard framework. *Industrial Management & Data Systems*, 106, 2, 242-255.
- Kaplan, R. S., & Norton, D. P. (Sept-Oct, 2000). Having trouble with your strategy? Then map it. *Harvard Business Review*, 78 No. 5. 1-11.
- Kaplan, R. S., & Norton, D. P. (2001). *The Strategy focused Organization: How Balanced Scorecard Companies thrive in the new business environment*. Harvard Business School Press.
- Kaplan, R. S., & Norton, D. P. (2001). Transforming the Balanced Scorecard from Performance Measurement to Strategic Management: Part I. *Accounting Horizons*, 15, 1, Retrieved from Questia Media America.
- Kaplan, R. S., & Norton, D. P. (2001). Transforming the Balanced Scorecard from Performance Measurement to Strategic Management: Part II. *Accounting Horizons*, 15, 2, Retrieved from Questia Media America.
- Kaplan, R. S., & Norton, D. P. (Feb 2004). Measuring the strategic readiness of intangible assets. *Harvard Business Review*, 1-14.
- Kaplan, R. S., & Norton, D. P. (2004). The strategy map: guide to aligning intangible assets, *Strategy and Leadership*. Harvard Business School Press, 32, 5, 10-17.

- Kaplan, R. S., & Norton, D. P. (2004). *Strategy Maps: Converting intangible assets into tangible outcomes*. Harvard Business School Press.
- Lester, Sr. J. D., & Lester, Jr. J. D. (2007). *Writing Research Papers: A Complete Guide*. 12th Ed. Pearson Education.
- Niven, P. R. (2005). *Balanced scorecard diagnostics: maintaining maximum performance*. John Wiley & Sons, Retrieved from Questia Media America.
- Norton, D. P. (1999). Use Strategy Maps to communicate your strategy. *Harvard Business School*.
- Norton, D. (1999). SAP Strategic Enterprise Management. Translating Strategy into Action: The Balanced Scorecard [Brochure].
- Othman, R. (2006). Balanced scorecard and causal model development: preliminary findings. *Management Decision*, 44, 5, 690-702.
- Papenhausen, C., & Einstein, W. (2006). Insights from the Balanced Scorecard Implementing the Balanced Scorecard at a college of business. *Measuring Business Excellence*, 10, 3, 15-22.
- Riordan, Matt. & Marcais, Tom. Moodle: An electronic classroom, Cuyahoga Valley Christian Academy & and Sweet Briar College Moodle Websites. 1.5.3, 1-127.
- Scholey, C. (2005). Strategy maps: a step by step guide to measuring, managing and communicating the plan. *Journal of Business Strategy*, 26, 3, 12-19.
- Thomas, H. (2007). Business school strategy and the metrics for success. *Journal of Management Development*, 26, 1, 33-42.
- Umashankar, V., & Dutta, K. (2007). Balanced scorecards in managing higher education institutions: an Indian perspective. *International Journal of Educational Management*, 21, 1, 54-67.
- Wibowo, Arrianto., & Yuwono, Budi. Patterns for describing IT value in IT strategy maps & IT balanced scorecard. IT Governance Lab, Faculty of Computer Science University of Indonesia press.
- Wilcox, M., & Bourne, M. (2003). Predicting Performance. *Management Decision*, Retrieved from Emerald Group.

Evaluating Energy Consumption Efficiency of the Zone-based Routing Protocol

Fahmy I.¹

Nassef L.²

Hefney H. A.¹

Abstract

In this paper, the hybrid routing protocol ZRP (Zone-based Routing Protocol) will be evaluated from an energetic perspective. ZRP divides a Mobile Ad-hoc wireless Networks (MANET) into a number of zones to reduce the routing overhead and distribute network traffic through a set of nodes to avoid frequent route recalculation. Therefore, ZRP is a protocol combining all the benefits of the "Proactive" and "Reactive" routing protocols and avoiding most of their drawbacks. In this paper, ZRP energy consumption efficiency evaluation as a critical QoS (Quality of Service) performance metric is mainly based on two main scenarios: The MANET size (number of nodes) and the nodes mobility represented by the maximum mobile node's speed and the pause time. The discrete event network simulation software QualNet was used to simulate these proposed evaluation scenarios.

Keywords: ZRP, MANET, Energy Consumption efficiency, QoS, and QualNet.

1. Introduction

The MANET represent a special category of wireless mobile networks that utilize multi-hop routing and are capable of operating without any fixed or wired infrastructure devices or equipments support [R. Shorey et al., 2006]. The MANET's major drawback is their difficult routing mechanism and overhead due to the absence of a base station during the packets transmission and reception among the mobile devices. MANETs have recently become suitable for different types of applications varying in their QoS priorities from home and road mapping to military and emergent ones due to their quick and easy deployment. Thus, such networks can be used in situations where either there is no other wireless communication infrastructure present or in situations where it is not possible to have such an infrastructure present due to an emergent situation (e.g. military battlefields or after natural disasters). They are also commonly used with collaborating and distributed computing, wireless mesh networks, wireless sensor networks and hybrid wireless networks [C. Murthy et al., 2004]. Figure 1 shows an example of the multi-hop routing in MANET to route a packet from source node A to destination node P through border node F and then border node M.

¹ Computer Sciences & Information Department, Institute of Statistical Studies & Research, Cairo University, Egypt
² Computer Science Department, Faculty of Computing & Information Technology, King AbdulAziz University, Jeddah, Kingdom of Saudi Arabia

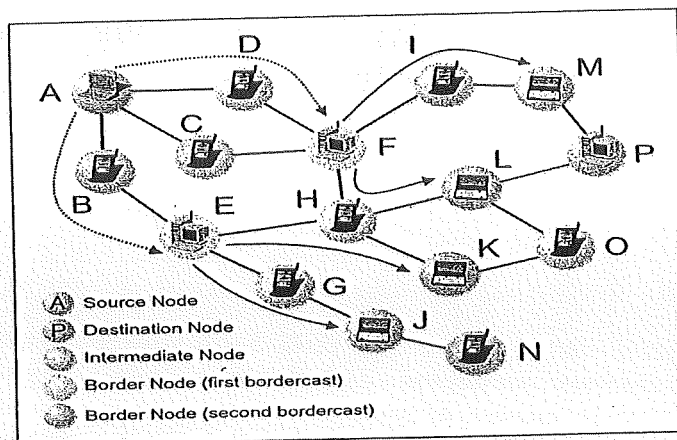


Figure 1: A multi-hop MANET example

In this paper, the ZRP components and operation are presented by the second section, then, the ZRP energy consumption efficiency and its standard model was discussed by the third section. In the fourth section, the simulation environment is described and then two main simulation scenarios: the MANET size and users mobility are highlighted. These scenarios are used to study the ZRP's behavior in terms of the energy consumption efficiency followed by their simulation results analysis in the fifth section.

2. Zone-based Routing Protocol

The Zone-based Routing Protocol (ZRP) components and its operation and zone creation are discussed briefly.

2.1 ZRP Components

ZRP is a hybrid routing protocol that divides the MANET into a number of overlapping zones. ZRP runs independent protocols within and between these zones. For intra-zone routing (within the same zone nodes) ZRP uses IARP (Intra-zone Routing Protocol) that works proactively and for inter-zone routing (between the nodes of different zones), it uses IERP (Inter-zone Routing Protocol) that works reactively on demand. The Border-cast Resolution Protocol (BRP) is also used to optimize the routing process between perimeter nodes [Z. J. Haas et al., 1999]. The topology of each routing zone (through a routing table or a network graph) is maintained proactively (IARP) and each zone member node is aware of the neighboring zones through the gateway nodes (BRP). The Neighbor Discovery Protocol (NDP) relies on the transmission of hello messages by each node. Even though the hybrid nature of the ZRP seems to indicate that it is a hierarchical routing protocol, but ZRP remains a flat protocol that has special privileges over the other flat routing protocols in MANET.

In figure 2, the ZRP components are presented to see how these three main protocols (Intra-zone, inter-zone, and border routing protocols) interact to provide an efficient flat routing protocol that acts proactively locally (within the same zone) and on-demand (reactively) out-zone.

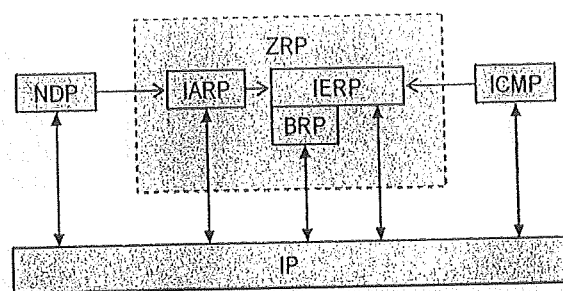


Figure 2: ZRP Components

The ZRP has numerous benefits in the infrastructure-less MANET environment as a hybrid routing protocol that is capable of adapting its behavior: It helps decreasing the routing overheads by reducing the search zone and allowing only selected nodes (zone-heads or cluster-heads) to forward the control packets. By grouping nodes together and allowing routing and data transmission to be carried out by a group of nodes, it eliminates single points of failure during data transmission, distribute network traffic through a set of nodes and avoid frequent route recalculation. Each node that migrates between zones also leaves transient zone trails to find the destination more quickly and with fewer overheads. Generally, ZRP significantly reduces the number of control packets transmitted into the network and achieves higher levels of packet delivery under worst case network conditions when compared to other MANET routing protocols. ZRP is a routing protocol that could be effective in larger size networks with applications that exhibit a relatively high degree of locality of communication, where communication between nodes with close proximity to one another is much more frequent than that between nodes which are further apart (out of zone).

2.2 ZRP Operation and Zone Creation

The hybrid routing protocol ZRP tends to divide the MANET into a certain number of zones [I. Fahmy et al., 2010]. These zones form each node's local neighborhood. Each node may be within multiple overlapping zones, and each zone may be of a different size. The size of a zone is not determined by a geographical area. It is rather determined by a radius of length, where the number of hops is the zone's perimeter. With the correct zone size, it is possible to reduce the control traffic to a minimum one. Each network configuration has an optimal zone radius value. To determine the optimal zone radius value, it is necessary to understand how different factors influence on the traffic amount [N. Bejar]. Each node has its own zone and can freely communicate with neighboring nodes through its zone's gateway nodes transparently [Z. J. Haas et al., 1999].

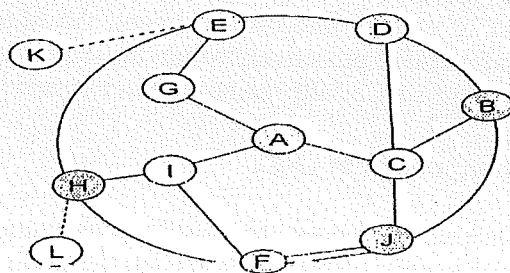


Figure 3: A ZRP MANET Configuration example

In figure 3, the MANET's zone radius is 2, and nodes: E, D, B, J, F and H are border-nodes. Before constructing a zone and determining its border nodes, each node should be aware of its local neighbors. It also requires a Neighbor Discovery Protocol (NDP) that relies on the transmission of hello messages by each node. For example, when node A gets a response from node B which has received the "Hello messages", node A notices that it has a direct point-to-point connection with B. The NDP selects nodes on various criteria, e.g., signal strength and frequency/delay of beacons. Once the local routing information has been collected, the node periodically broadcasts discovery messages in order to keep its map of neighbors up to date. Sometimes the MAC layer of the nodes does not allow for such a NDP. Then the Intra-zone Routing Protocol must provide the possibility of direct neighbor discovery. This protocol is responsible for determining the routes to the peripheral nodes and is commonly a proactive protocol.

3. ZRP Energy Consumption Efficiency

The node's limited rechargeable battery power as an energy resource in MANET should be considered as a critical QoS (Quality of Service) performance metric that could severely affect the entire network if not preserved and consumed efficiently during transmission.

3.1 ZRP Energy Consumption

QoS provisioning for wireless MANETs is a critically challenging issue due to the self organizing nature of such networks, and also due to the strict energy constraints and signaling overhead. Therefore, the QoS new perspective adopted by this paper and its new performance metrics should differ from other traditional wireless networks performance metrics.

With the node energy constraints added by limited battery power as the MANET users are moving most of the time with their mobile nodes/terminals, power control and energy management can greatly improve the network performance. On the other hand, the network performance will depend on the varying network topology which is influenced by routing [H. V. Peeter et al.]. Therefore, QoS in MANET is defined as the performance level of a service offered by the network to the user [P. Sholander et al.] to achieve a more deterministic network behavior, implying that information carried by the network can be better delivered by controlling the throughput and delay time. And meanwhile, the network resources can be efficiently utilized by energy consumption management since the node's energy resource is considered a critically scarce resource in MANET that should be preserved and efficiently utilized.

ZRP protocol, as a hybrid routing protocol dividing the MANET into overlapping zones, helps the mobile nodes to preserve their battery power while communicating and moving. The distribution of control information among the MANET zones and the maintenance of control information proactively within the same zone enable an efficient use of all the MANET resources especially the nodes memory and battery power.

3.2 Node Energy Consumption in MANET

The performance of MANET has been analyzed from an energetic point of view by [M. Fotino et al.]. In their energy consumption model, the generic expression used to calculate the energy required to transmit a packet p : $E(p)$ in joules is given by (1):

$$E(p) = i * v * t_p \quad (1)$$

Where i represent the current consumption, v is the voltage used and t_p is the required time in seconds to transmit a packet given by (2):

$$t_p = \left(\frac{p_h}{6 * 10^6} + \frac{p_d}{54 * 10^6} \right) \quad (2)$$

Where p_h is the packet header size and p_d is the packet data size (both in bits). Then, the energy consumed by the node in transmit mode $E_t(p)$ is given by (3), while the energy consumed in reception mode $E_r(p)$ or in overhearing mode when the node overhears the packets exchanged within its range are given by (4):

$$E_t(p) = 280mA * v * t_p \quad (3)$$

$$E_r(p) = E_o(p) = 240mA * v * t_p \quad (4)$$

Therefore, the total amount of energy consumed at a node n_i is calculated by (5):

$$E(n_i) = E_t(p_{n_i}) + E_r(p_{n_i}) + E_o(p_{n_i}) \quad (5)$$

Since the energy consumed during transmission or reception is related to the communication process, it could be considered as efficient energy consumption. However, the energy expenditure, while overhearing the packets exchanged or at idle mode, is exhausting the node's battery inefficiently. Therefore, the energy consumed at idle mode and while overhearing packets exchanged should be minimized by determining a smaller "idle" time after which the node automatically goes to "sleep" mode to save its battery. Moreover, as the number of nodes increases in a MANET, a node's neighbors increase as well resulting in more energy consumed in overhearing.

In this paper, the energy consumption and hence efficiency for the hybrid routing protocol ZRP in MANET will be evaluated to determine whether it has improved the QoS perceived by a data session, where the measure of data QoS is taken to be the energy required for the transmission of one bit. Therefore, the MANET's hybrid routing protocol ZRP energetic behavior was inspected in order to be an energy efficient base routing protocol to measure the energy efficiency degree of other routing protocols under research.

Although a MANET using the hybrid ZRP can profit of both proactive and reactive routing protocols benefits, its performance response, when the network density increases, or when the maximum speed of the mobile nodes or their pause time (in mobility model) changes, should be carefully observed in order to evaluate it in terms of the most challenging QoS evaluation metric: The energy consumption at transmit mode $E_t(p)$ calculated as the value of energy consumed by nodes in Joules while transmitting and at receive mode $E_r(p)$ calculated as the value of energy consumed by nodes in Joules while receiving.

4. ZRP Simulation Scenarios

In this section, the "QualNet" network simulation tool introduced by Scalable Networks [Scalable Networks, 2008] is actually used to simulate the MANET wireless environment employing ZRP for two main scenarios to evaluate its performance [C. N. Ververidis et al.] according to the energy consumption QoS metrics previously mentioned. The simulation is mainly based on the two following scenarios.

4.1 Scenario 1: The MANET Size

In this scenario, the impact of the MANET size is depicted in terms of the energy consumption at transmit and at the receive modes as the number of mobile nodes increases from 50 to 100 nodes.

4.2 Scenario 2: The Users Mobility

In this scenario, the impact of users mobility, represented in the first sub-scenario by the mobile node's pause time and in the second sub-scenario by the mobile node's maximum speed, on the amount of energy consumed.

4.2.1 Varying the mobile node's "Pause Time":

In this sub-scenario, the impact of varying the pause time of the mobile node for studying its behavior (30 sec was the default value) as its pause time decreases (15 sec) or increases (45 sec) for a medium size MANET with 75 nodes.

4.2.2 Varying the mobile node's "Maximum Speed":

In this sub-scenario, the impact of varying the maximum speed (10 m/sec was the default value) of the mobile node for studying its behavior as its maximum speed decreases (5 m/sec) or increases (20 m/sec) for a medium size MANET with 75 nodes. Generally, these scenarios, described earlier, have the following simulation environment parameters as shown by table 1.

Table 1: The Simulation parameters

Configuration parameter	Parameter value and unit
Cartesian coordinate system terrain area	1500*1500 meters
Simulation time	30 seconds
Experiment seed	1
Channel frequency	2.4 GHz
Propagation limit	-111.0 dBm
Noise factor	10.0
Data Rate	2 Mbps
Transmission power	15.0
Nodes Placement	Random
Path-loss model	Free-space
Application traffic	CBR

5. Simulation Results and Analysis

The simulation results and their analysis are described by the following sub-sections.

5.1 Scenario 1: Impact of increasing MANET size on energy consumption

In figure 4, it is clearly shown that as the number of mobile nodes increases, the energy consumption at the transmit mode increases and varies according to the users mobility behavior which causes the graph's overlapping (blurring) behavior.

Figure 5, the graph is showing a definitely lower energy consumption in the receive mode at higher number of mobile nodes (100 nodes) implying that as the MANET size increases the amount of energy consumption by the entire network decreases which helps saving the energy consumed as a scarce resource specially in the limited rechargeable battery power nodes MANETs.

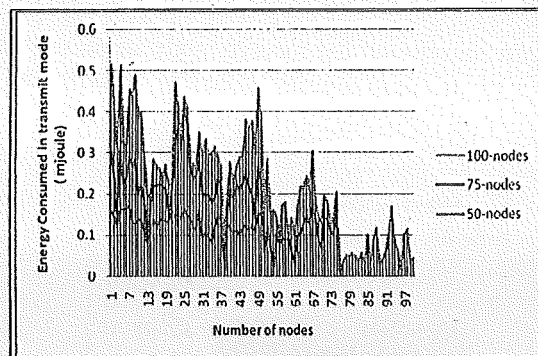


Figure 4: The impact of MANET size on energy consumed in transmit mode

On the other hand, at the receive mode, as inspected by figure 5, the energy consumed at each MANET size keeps almost a constant value in milli-Joules. As the network number of nodes increases to a medium size, the energy consumed in receive mode is definitely higher (directly proportional) but it dramatically degrades at larger sizes but maintaining a relatively higher value of energy consumed than the smaller MANET sizes.

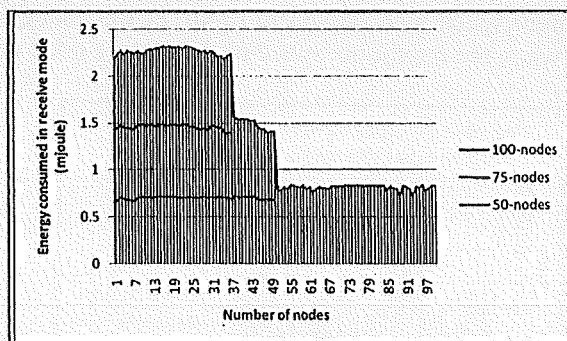


Figure 5: The impact of MANET size on energy consumed in receive mode

5.2 Scenario 2: Impact of users mobility on energy consumption

5.2.1 Impact of varying the "Pause time" on energy consumption

In figure 6, the energy consumption in transmit mode is directly proportional to the mobile node's pause time; as the mobile node's pause time increases, the energy consumed in mJoules increases proportionally.

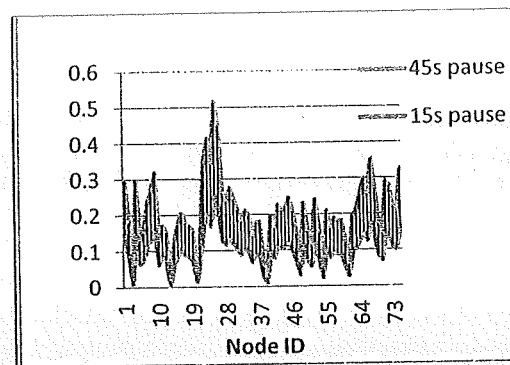


Figure 6: The impact of varying the pause time on energy consumed in transmit mode

In figure 7, the energy consumption at the receive mode increases as the pause time increases proportionally but in an approximately constant behavior alike the transmit mode.

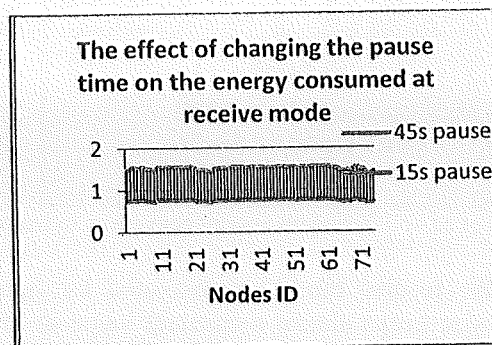


Figure 7: The impact of varying the pause time on energy consumed in receive mode

5.2.2 Impact of varying the "Maximum speed" on the energy consumption

The energy consumption in mJoules at transmit mode is inversely proportional to the mobile terminal's maximum speed as observed in figure 8; the energy consumption at transmit mode increases when the maximum speed of the mobile node decreases.

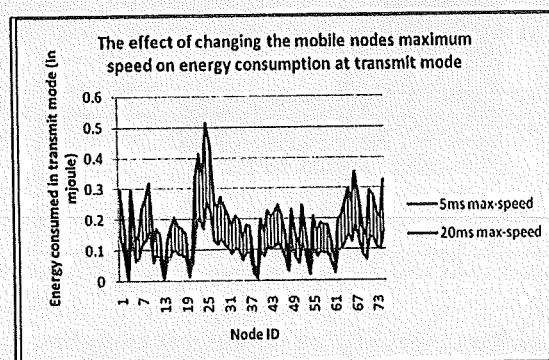


Figure 8: The impact of varying the maximum speed on energy consumed at transmit mode.

The energy consumption in mJoules in receive mode is also inversely proportional to the mobile node's maximum speed as shown by figure 9; the energy consumption amount in receive mode increases when the maximum speed of the mobile terminal decreases.

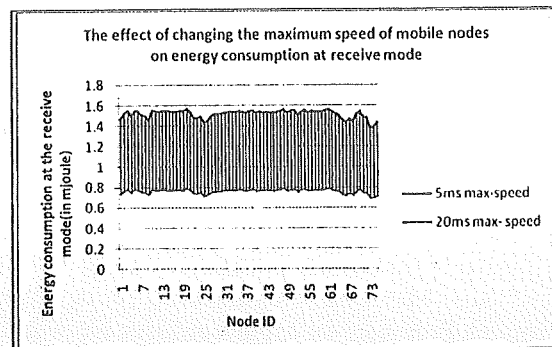


Figure 9: The impact of varying the maximum speed on energy consumption at receive mode.

6. Conclusion and Future Work

In this paper, the Mobile Ad-hoc wireless Networks, commonly referred to as MANET, were briefly introduced, then the hybrid routing protocol ZRP (Zone-based Routing Protocol) components, operation and zone creation were discussed. Then, the paper has highlighted the relationship between QoS provisioning in MANET and the proposed energy consumption and efficiency evaluation metric.

Then, ZRP was evaluated according to two main scenarios; the first was the MANET size, the second was the user mobility represented by the node's pause time and its maximum speed.

Finally, it could be deduced from the simulation statistics analysis demonstrated by the previous section that the MANET's hybrid routing protocol ZRP is maintaining satisfactory and efficient energy consumption at relatively large MANET size with moderate user's mobility concerning pause time and maximum speed values. Also, as the user mobility varies in terms of node pause time or maximum mobility speed, the ZRP maintains a relatively lower energy consumption at both transmit and receive modes.

Therefore, the hybrid zone-based routing protocol ZRP could be considered as a reference base protocol from the energetic point of view for the future research work concerning the energy model optimization in MANET.

References

[C. N. Ververidis et al.]

Christopher N. Ververidis and George C. Polyzos, "A Routing Layer Based Approach for Energy Efficient Service Discovery in Mobile Ad Hoc Networks", Mobile Multimedia Laboratory, Department of Computer Science, Athens University of Economics and Business, 47A Evelpidon & Lefkados str., Athens 11362, Greece.

[C. S. R. Murthy et al., 2004]

C. Siva Ram Murthy, B. S. Manoj, "Ad Hoc Wireless Networks Architectures and Protocols.", Prentice Hall Professional technical reference, New Jersey, 2004.

[H. V. Peeter et al.]

H. Vincent Poor, Cristina Comaniciu, "QoS provisioning for wireless Ad-hoc Data networks." Department. of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544.

[I. Fahmy et al., 2010]

Imane M. A. Fahmy, Laila Nassef, Imane A. Saroit, and S. H. Ahmed, "QoS Parameters Improvement for the Hybrid Zone-Based Routing Protocol in MANET", INFOS 2010, Cairo university, March 2010.

[M. Fotino et al.]

Marco Fotino, Antonio Gozzi, Floriano De Rango, Salvatore Marano, Juan-Carlos Cano, Carlos Calafate, Pietro Manzoni, «Evaluating Energy-aware Behavior of Proactive and Reactive Routing Protocols for Mobile Ad Hoc Networks»

[N. Beijar]

Nicklas Beijar, "Zone Routing Protocol (ZRP)", Networking Laboratory, Helsinki University of Technology P.O. Box 3000, FIN-02015 HUT, Finland.

[P. Sholander et al.]

Peter Sholander, Paul Coccoli, Tracey Oakes and Sean Swank, Scientific Research Corporation, "A portable Software Implementation of a hybrid MANET Routing Protocol.", Atlanta, GA, U.S.A.

[Scalable Network, 2008]

Scalable Network Technologies, "QualNet-4.5 Wireless-Model Library", www.scalable-networks.com, March 2008.

[R. Shorey et al., 2006]

Rajeev Shorey, A. Ananda, Mun Choon Chan, Wei Tsang Ooi, "Mobile, Wireless and Sensor networks, technology, applications, and future directions. ", A. John Wiley & Sons publication, WILEY INTERSCIENCE, IEEE PRESS, 2006.

[Z. J. Haas, 1999]

Zygmunt J. Haas, "A New Routing Protocol for The Reconfigurable Wireless Networks.", School of Electrical Engineering, Cornell University Ithaca, NY 14853, 1999.

Towards Building Annotated Arabic Web-based Corpus

Ezzat M. *

ElGhazaly T. **

Abstract

The Web is a potentially unlimited source of linguistic data; however, commercial search engines are not the best way for linguists to gather data from it. Here in this paper, we present a new methodology for building language corpora, in the first section we provide an introduction on the corpus, in the second section we describe the general applications of the corpus and in the third section we describe three approaches used for building web corpus.

In section four we describe our new methodology which includes Crawl seeding and crawling, where we start crawling based on predefined seeds of word lists and URLs, Post-crawl cleaning, which describes the procedures we propose to clean the results we got from by the crawling. And finally Part-of-speech tagging/lemmatization and indexing where we annotate the cleaned results we got from the first two steps.

Keywords

Corpora, Corpus, Linguistic corpus, Web Corpus, Arabic Corpus.

1. Introduction

A number of large corpora have been assembled in the last few years, but the idea itself is not new. It can be traced back to the German linguist Kading, who in 1897 used a large corpus of German - 11 million words - to collate frequency distributions of letters and sequences of letters. This corpus, by size alone, is impressive for its time, and compares favorably with more modern corpora (McEnery and Wilson, 1996). In the late 20th century, there were attempts to develop a number of large corpora of language which can be electronically searched. Early signs of the modern era of corpus linguistics are dated by McEnery and Wilson (1996) from 1960 when Quirk planned and implemented his ambitious Survey of English Usage (SEU). At the same time, Francis and Kucera began compiling the Brown corpus, which was developed over the following two decades. In 1975 Jan Svartvik began extending the work of the SEU and the Brown corpus to construct the London-Lund corpus. In 1994 the Cambridge and Nottingham Corpus of Discourse in English (Cancode) began, funded by Cambridge University Press and presently contains 5 million words of transcribed data.

1.2. What is a Corpus?

Corpus, plural corpora a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language. For example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study.

* Department of Computer Sciences, Institute of Statistical Studies and Research, Cairo University, Egypt

** Department of Computer Sciences, Institute of Statistical Studies and Research, Cairo University, Egypt.

2. Applications of corpora in language research

The importance of corpora to language and linguistics studies is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalized cognitive perception of language. As language and linguistics studies cannot rely on intuition or small samples of language; they require empirical analysis of large database of texts as in the corpus-based approach. Corpus-based methods can be used to study a wide variety of topics within linguistics. Because Corpora consist of texts, they enable the linguists to contextualize their analyses of language; corpora are very well suited to more functionally based discussions of language and linguistics. Modern computers have made it possible to store a large number of texts and to analyze a large number of linguistic features in those texts.

2.1 Why we need Arabic Corpora?

Collecting manuscripts, books and newspapers for analysis is very laborious in nature. But this was done for a long time, particularly by Academic researchers. Thankfully, as technological advances make the computerized storage of and access to large quantities of information easier, so the construction and use of text corpora will continue to increase. As a result the potential for research has widened considerably [2].

Attempts to study Arabic using these types of resources were initiated by researchers in the NLP field. To evaluate Information Retrieval systems and morphological analyzers Al-Kharashi & Evens in 1994 [3] used a collection of bibliographical records to test their Information Retrieval Micro-AIRS. Hmeidi et al. constructed a corpus of 242 abstracts collected from the proceedings of the Saudi Arabian national conference [4]. Xu, Fraser, and Weischedel [5] collected a parallel corpus from the United Nations (UN) published documents. The United Nations web site (<http://www.un.org>) publishes all UN official documents under a document repository, which is accessible by paying a monthly fee. A special purpose crawler was used to extract documents that have versions in English and Arabic. After a series of cleanups, they obtained 38,000 document pairs with over 50 million English words. The corpus was used to solve the issues of broken plural, which generally are not handled by stemmers.

2.3. Advantages and drawbacks of using the Web as a corpus

One of the main advantages of constructing Web-based corpora is size. Size matters in NLP. In an influential paper [6], Banko and Brill have shown that even a simple disambiguation algorithm outperforms more sophisticated methods when it is trained on a larger amount of data, and that, even after seeing a training corpus of one billion words, the size-dependent increase in performance does not show signs of having approached an asymptote. However, size matters also in other more theoretical fields of linguistics. For example, Mair [7] has shown that theWeb, unlike the British National Corpus (BNC, [8]) BNC, is large enough to allow a full study of the grammaticalization of get as a passive in English. Turney [9] showed how a simple algorithm relying on Web frequencies outperforms a much more sophisticated method trained on a smaller corpus in a synonym detection task. For very specialized sub-languages, the vastness of the Web might be crucial in another sense: even if we intend to build a small corpus, only a database as large as the Web will provide enough documents to be included in such corpus [10].

The second advantage of theWeb is that it allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to

The second advantage of theWeb is that it allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers. This set does not include only so-called "minority languages", but also well studied languages such as Arabic. The results of [11] and [12] suggest that Web corpora built by a single researcher literally in minutes are, in terms of variety of genres, topics and lexicon represented, closer to traditional "balanced" corpora such as the BNC than to mono-source corpora, such as newswire-based corpora. Moreover, these corpora will tend to reflect more recent phases of a language than traditional corpora that are often subject to a certain lag between the time of production of the materials that end up in the corpus and the publication of the corpus itself.

The third advantage of Web corpora is that they can potentially contain a number of genres that are not present in traditional written sources. Phenomena such as blogging should be of interest to linguists since they generate vast amounts of written samples on a huge variety of topics that are spontaneously produced by non-professional writers. Furthermore, the Web provides plenty of samples of interactive communication that, while in written form, possesses some characteristics of oral communication [25]. At the same time, as usage of the Web for various archival and practical purposes spreads, it is hard to think of traditional written genres that are not represented online.

Web corpora, of course, pose also some problems. First, such corpora tend to contain a lot of noise, such as automatically generated non-linguistic material and duplicated documents. Second, perhaps more worryingly, since Web corpora are typically constructed with automated text mining methods, the researcher often does not have full control over what ends up in the corpus, and can only estimate the composition of the corpus with post-hoc methods. As a consequence of this, the corpus is likely to be defective in terms of meta-data information. Third, if a researcher plans to distribute a largeWeb corpus made of millions of documents, (s) he will have a very hard time obtaining permission to use the documents from all the copyright holders. Notice that none of these issues are unique to Web corpora. Rather, they come to the forefront with Web corpora because, using Web mining methods, one can collect a very large corpus in a very short time. If one collected a Web corpus of about 100M words spending the same amount of time and resources that were invested in the creation of the BNC, there is no reason to think that the resulting corpus would be less clean, its contents less controlled or its copyright status less clear than in the case of the BNC. Vice versa, collecting a one billion word multi-source corpus from non-Web sources in a few days is probably not possible, but, if it were possible, the resulting corpus would almost certainly have exactly the same problems of noise, control over the corpus contents and copyright that we listed above. Thus, we would like to stress that it is not correct to refer to the problems above as "Problems of Web corpora"; rather, they are problems of large corpora built in short time and with little resources, and they emerge clearly with Web corpora since the Web makes it possible to build "quick and dirty" large corpora. It is a matter of research policy, time constraints and funding to determine if, for a certain project, it is better to invest considerable time and funds in building a thoroughly controlled, probably relatively small corpus, or if it is better (or: the only viable solution given external constraints) to build a possibly very large corpus that will probably suffer of all the problems above.

The very fact that it is relatively easy, very fast and cheap to construct Web corpora might actually provide innovative solutions to some of the issues described above. In particular, for certain lines of research it might be sufficient to share rapid corpus construction methods, rather than the data themselves, thus overcoming the copyright problems: a scientist could replicate the experiment of another scientist not on the same corpus, but on a corpus constructed according to the same criteria. This would be equivalent to what is commonly done in experimental fields of linguistics, such as phonetics, where scientists in different labs

replicate experiments on comparable samples of subjects – not on the very same subjects. While, in general, we do not see any cons that are unique to Web corpora, there are, nonetheless, problems specific to particular Web-as-corpus methodologies, namely those that rely heavily on Google or other commercial search engines to obtain Web data. We turn now to a discussion of this and other approaches to using the Web as a corpus.

3. Three approaches to the Web as a linguistic corpus

3.1. Hit counts as frequency estimates

Probably the oldest and most widespread approach to using the Web for linguistic purposes is to issue a query to a search engine such as Google, and to use the “hit count” reported by the search engine as an estimate of the frequency of occurrence of the searched string in the target language. This strategy has proven very successful in various tasks. To quote just one famous example, Turney [27] showed that a simple approach based on collecting hit counts of queries of the type A NEAR B to the AltaVista search engine performed better, in a synonym detection task, than a much more sophisticated method (latent semantic analysis) that used a traditional corpus as input.

However, despite its empirical successes, this approach is very problematic. First, the types of queries that one can issue to a search engine are very limited: for example, one cannot restrict the search on the basis of parts-of-speech, and one cannot use regular expressions. Indeed, since search engine users are typically interested in what their search terms refer to, rather than on their linguistic properties, search engines tend to perform a number of normalizations of the search terms that can be extremely annoying for linguists, such as ignoring case, stripping off diacritics, ignoring apostrophes and dashes. For similar reasons, search engines often ignore or produce very strange results when queried for function words. Thus, the types of research questions that one can approach with this methodology are a priori very limited: one can use search engine counts to look for the frequency of preselected content words and word combinations, and only if exact matching is not crucial to the research question.

Second, search companies, for obvious reasons, do not publish detailed information on how they gather, index and return query results, and the services they provide, being often and unpredictably updated following technological and market changes, tend to be extremely brittle. A very dramatic illustration of this problem took place in April 2004, when AltaVista suddenly stopped supporting the NEAR operator, making Turney’s original algorithm unusable. More recently, Jean V’eronis has been reporting on his blog⁵ a series of experiments pointing out various ways in which the counts of pretty much all the major search engines are unreliable and often obviously inconsistent. In short, using search engine hits seems only appropriate for pilot studies, or in very restricted contexts, but it is not a viable long-term approach to using the Web as a corpus.

3.2. Building corpora through search engine queries

Rather than relying on search engine counts, one can issue automated queries to the search engine (search engines such as Google and Yahoo! provide Web service APIs that allow users to perform a certain number of automated queries per day), retrieve the pages returned by the search engine, and process them to build a corpus.

This approach has been explored by various scholars, including [15, 10, 14], and very extensively by Sharoff [11] and Ueyama [12]. Sharoff shows how corpora in various languages built by issuing queries for random combinations of frequent words to the Google

search engine and retrieving and processing the pages found in this way have characteristics more similar to those of a balanced corpus like the BNC than to those of mono-source corpora. Ueyama applies the same methodology to Japanese, and shows, in particular, how repeating the procedure at different times produces corpora that are dramatically different in terms of the pages retrieved, but relatively comparable in terms of their qualitative makeup (distribution of genres and topics). On the other hand, different choices of query terms lead to corpora with very different make-ups. In particular, using basic vocabulary words as query terms leads to corpora that are characterized by a high proportion of personal, everyday life prose produced by non-professional writers, whereas query terms sampled from more formal sources lead to corpora that are characterized by public/scientific/technical topics and professional authors.

This approach is less heavily reliant on Google (or other search engines) than the one discussed in the previous section. Google is used to obtain a list of documents, but then these documents are retrieved and post-processed by the researcher (tokenized, POS-tagged, etc.) locally, so that the stability of the data will no longer depend on Google, the researcher has full access to the corpus and, with the appropriate tools, the corpus can be interrogated with sophisticated linguistic queries.

However, the approach is not devoid of problems. The set of pages that are retrieved is still dependent on the search engine matching and ranking criteria. Moreover, for obvious reasons search engines restrict the amounts of data that can be obtained by automated querying (e.g., in the case of Google one can maximally retrieve 10K result URLs per day – and, of course, not all the pages retrieved are appropriate for a corpus).

Thus, while the vast amount of data available there is one of the main factors attracting linguists to the use of the Web, building truly large corpora (in the order of millions of documents) with this method is extremely impractical. More in general, providing a legal automated query interface is unlikely to be a high priority of search engines, and it would not be too surprising if the search engine companies stopped supporting such services, or started charging for them. Thus, it is not clear that even this is a viable long term approach.

3.3 Linguistic crawls of the Web

We believe that the only viable long term approach to constructing Web corpora is for linguists to perform their own crawls of the Internet. This makes linguists fully independent from commercial search engines, and provides full control over the whole corpus construction procedure. However this is also the most difficult approach to implement, especially if the target is a large corpus. Considerable computational resources are necessary to host a large scale crawl; the data produced by the crawl have to be “cleaned” (removing pages not in the target language or problematic for other reasons; stripping off html code and “boilerplate”, discarding duplicates).

The data must then be annotated, minimally, with part of speech tags and lemmas (adapting the annotation tools to the language of the Web). When the input data are in the order of hundreds of gigabytes, all these steps become far from trivial, not only from the point of view of linguistic quality of the results, but also in terms of time and efficiency.

Indeed, as far as we can tell, none of the previous projects of large linguistic crawls of the Web went through all the stages of the outlined procedure. The terabyte corpus described in [16] has not undergone language filtering nor other forms of post-processing, and it is not annotated with linguistic information. The English Academic Web site corpus of [17] has also not undergone any form of post-processing, besides simple tokenization. The Chinese corpus of [18] has been filtered in terms of language detection, but it is not processed in any other way.

We believe that the kind of resources and know-how necessary to build, annotate and maintain large Web corpora will be attained more easily by a community of linguists interested in this task.

4- Constructing a large general purpose Web-based corpus for Arabic

In this section we will describe the steps to be followed for building the corpus, these steps includes crawl the web, cleaning the results and then putting the part of speech tagging, and the following subsections will describe each step in more detail.

4.1 Crawl seeding and crawling

We suggest that the crawl must start from a set of pre-selected URLs (the crawl "seeds"). Obtained from a list of URLs in Arabic by random queries to Google for combinations of frequent Arabic words, taken both from a basic vocabulary list and from a corpus of articles from the Arabic newspapers, and retrieving maximally 10 pages per query. The crawl of pages will be performed by a spider with a multi-threaded breadth-first crawling strategy, avoiding pages that's URLs end in a suffix cuing non-html data (.pdf, .jpeg, etc.)

4.2. Post-crawl cleaning

In this section we will describe the steps we suggest to clean what we got from the crawling, we are using these steps because we know that the results of the crawling will contain many unneeded text and

4.2.1. First pass filtering

Before further processing, we shall discard all retrieved documents that are not of mime type text/html, and documents smaller than 5KB or larger than 200KB. We also shall remove all sets of documents with perfect duplicates in the collection.

4.2.2. Code removal and boilerplate stripping

Besides html and java script code, Web-pages often contain link lists, navigational information, fixed notices, and other sections poor in human-produced connected text. From the point of view of corpus construction, boilerplate identification is critical: Too much boilerplate will invalidate statistics collected from the corpus and will impair attempts to analyze the text by looking at keywords in context. Boilerplate is much harder to spot than code that is relatively easy to identify with regular expressions.

Based on the idea that the content-rich section of a page will have a low html tag density, whereas boilerplate text tends to be accompanied by a wealth of html (because of special formatting, many newlines, many links, etc.). Thus, among all possible spans of text in a document, we pick the one for which the quantity $N(\text{tokens}) - N(\text{tags})$ takes the highest value. Then we can remove html tags after they are used for the count.

The boilerplate stripping method used is based on general properties of Web documents, and thus relatively independent of the language and crawling strategy. Moreover, the method is not demanding on memory and it can be run in parallel on different machines, since it analyzes one document at a time. This is not the case with other approaches we considered,

which are based on searching for frequent strings in the whole collection. On the negative side, the method tends to return a content-rich fragment of each page. As such, the resulting corpus is only appropriate for those who are interested in large collections of unstructured samples of natural language, where as it does not contain reliable data to study the structure of Web documents.

4.2.3. Near-duplicate detection

Next, we shall look for near duplicates, that is, documents that (while not identical) contain substantial overlapping portions. We can use a simplified version of the “shingling” algorithm [19], implemented in Perl/MySQL. For each document, after removing all function words, we take fingerprints of a fixed number s of randomly selected n -grams (sequences of n words; we count types, not tokens – i.e., we only look at distinct n -grams, and we do not take repetitions of the same n -gram into account); then, for each pair of documents, we count the number of shared n -grams, which can be seen as an unbiased estimate of the overlap between the two documents. We look for pairs of documents sharing more than t n -grams, and we discard one of the two. The pairs are ordered by document ID. To avoid inconsistencies, we can always remove the second document of each pair. Thus, if the pairs A-B, B-C and C-D are in the list, only the document A is kept; however, if the list contains the pairs A-C and B-C, only C is removed. We leave it to further research to devise efficient ways to identify clusters of near-duplicates.

More precisely, we extract 25 5-grams from each document, and we treat as near-duplicates documents that share at least two of these 5-grams. This threshold might sound surprisingly low, but the chances that, after boilerplate stripping, two unrelated documents will share two sequences of five content words are very low. Near-duplicate spotting took about one day, and it resulted in a corpus containing 1,875,337 documents, for a total of about 10GB of uncompressed data (notice the dramatic shrinking from the 81GB of compressed data we initially obtained from the crawl!).

4.3. Part-of-speech tagging/lemmatization and indexing

Then we can perform part-of-speech tagging with the widely used Tree Tagger, and generate an annotated corpus generated from the web in order to use the corpus in an efficient and linguistically sophisticated way, indexing needs to be carried out with a program that allows fast retrieval and sophisticated queries, for these purposes, we use the any popular free indexing and retrieval toolkit for very large corpora.

5. Conclusion

In this paper, we discussed the potential advantages of Web data for linguistic work, and we argued that such data should be accessed by crawling the Web directly, rather than using commercial search engines. We then illustrated the expected work necessary to build Web corpora.

We also presented short examples of the sort of linguistic analysis that Web data are good for, either because of size or because of the genres they represent. While we believe that there is no major theoretical or computational roadblock to the development of large Web-based

corpora, our current research shows that there are many practical issues to be dealt with, if we want to build high quality Web corpora.

To conclude on an optimistic note, however, we believe that our preliminary assumption shows that even the current Web corpora, despite all their noise and indexing issues, can be very useful in various areas of linguistic research.

References

- [1](cf. McArthur, Tom "Corpus" , in: McArthur, Tom (ed.) 1992. *The Oxford Companion to the English Language*. Oxford, 265-266)
- [2] Hunston, S. "Corpora in applied linguistics," Cambridge University Press May 2002.
- [3] Al-Kharashi, I. A. and Evans, M. W. "Comparing words, stems, and roots as index terms in an Arabic information retrieval system," *Journal of the American Society for Information Science (JASIS)* 5(8), pp 548-560, 1994.
- [4] Abdelali, A., Cowie, J., Soliman. H. "Building A Modern Standard Arabic Corpus," Workshop on Computational Modeling of Lexical Acquisition, the Split Meeting, Croatia, 25th to 28th of July 2005.
- [5] Xu, J. Fraser, A. Weischedel M. R. (2001) TREC 2001 Cross-lingual Retrieval at BBN NIST Text RE-trieval Conference TREC10 Proceedings, Gaithersburg, MD, pp.68-77.
- [6]Banko, M., and Brill, E. (2001) 'Scaling to very large corpora for natural language disambiguation', *Proceedings of ACL-01*.
- [7] Mair, Ch. (2003) 'Tracking ongoing grammatical change and recent diversification in present-day Standard English: the complementary role of small and large corpora', paper presented at the Annual ICAME Conference.
- [8] Aston, G., and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- [9] Turney, P. (2001) 'Mining the Web for synonyms: PMIIR versus LSA on TOEFL'. *Proceedings of ECML 2001*, 491-502.
- [10] Baroni, M., and Bernardini, S. (2004) 'BootCaT: Bootstrapping corpora and terms from the web', *Proceedings of the Fourth Language Resources and Evaluation Conference*.
- [11] Sharoff, S. (2006) 'Creating general-purpose corpora using automated search engine queries', in [13].
- [12] Ueyama, M. (2006) 'Creation of general-purpose Japanese Web corpora with different search engine query strategies', in [13].
- [13] Baroni, M., and Bernardini, S. (eds.) (2006) *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit.
- [14] Fletcher, B. (2004) 'Making the Web more useful as a source for linguistic corpora'. In *Corpus Linguistics in North America 2002*, U. Connor and T. Upton (eds.) Amsterdam: Rodopi.
- [15] Ghani, R., Jones, R., Mladenic, D. (2001) 'Using the Web to create minority language corpora'. *Proceedings of the 10th International Conference on Information and Knowledge Management*.

- [16] Clarke, C., Cormack, G., Laszlo, M., Lynam, T., and Terra, E. (2002) 'The impact of corpus size on question answering performance.' Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval.
- [17] Thelwall, M. (2005) Creating and using Web corpora. *International Journal of Corpus Linguistics* 10(4), 517-541.
- [18] Emerson, T., O'Neil, J. (2006) 'Experience building a large corpus for Chinese lexicon construction'.
- [19] Broder, A., Glassman, S., Manasse, M., Zweig, G. (1997) 'Syntactic clustering of the Web'. Proceedings of the Sixth International World-Wide Web Conference.
- [20] M. Baroni, and M. Ueyama (2006) 'Building General- and Special-Purpose Corpora by Web Crawling' Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application .
- [21] Tarek Elghazaly. "Cross Language Information Retrieval for Digital Libraries with Arabic OCR-Degraded Text" A PhD Dissertation, Cairo University, Faculty of Computers & Information, 2010.

An Enhanced framework for Measuring ERP Implementation Success In the Egyptian Environment

Esmat M. Abdelmoniem¹

Sherif A. Mazen²

Ehab E. Hassanein³

ABSTRACT:

Enterprise Resources Planning (ERP) selection process is an important step which impact on the success of any organization. Empirical research has shown that selecting an inappropriate system is a major reason for ERP implementation failures, in Egypt the success rate of implementing ERP systems is extremely lower than that in western companies [1].

In this paper we adopted a framework [2]. For ERP implementation in Egypt, including the factors that guarantee ERP systems implementation success. This framework covers 21 factors that can be divided into 4 categories (Top management, Project, Organization, and System) to assess the influence of Critical Success Factors (CSFs) on the implementation of ERP systems on large companies in Egypt. Data collection is done through an extensive review and many interviews with IT managers in some Egyptian enterprises using ERP system. The researchers have designed the questionnaires that measure the Factor to expedite the process.

KEYWORDS:

Enterprise Resource Planning (ERP), ERP success and failure.

1. INTRODUCTION

In the literature, it was reported that 75% of the ERP projects are classified as failures [3], 51% viewed their ERP implementation as not fully completed. Based on the ERP survey conducted by Meta Group in 2002, [4] the average cost of ERP ownership were \$15 millions ranging from half millions to \$300 millions. The average cost per user per year could be as high as \$20,000. However, in other frequent reports of ERP failure: "between 50% and 75% of U.S. experienced firms some degree of failure. One recent survey [1] revealed that 65% of executives believe that ERP implementation has at least a moderate chance of hurting their business. "Three quarters of the ERP projects are considered failures and many ERP projects ended catastrophically"

2. ERP BENEFITS AND OUTCOMES

- Reducing time and cost associated with warranty and the returns of processing [5].
- Improving communication and coordination among cross-functional teams.
- Improving visibility into capacity, inventory, and labor.
- Reducing in delivery time and transportation costs through vehicle-routing analysis.
- Determining of weights of influence over functioning organizations in developing countries [6] (like Egypt).
- Determining the impact of successful ERP implementation on the organizations [6].
- A successful ERP system implementation is considered to be implemented within budget, on time and delivers anticipated functionalities [7].

3. ERP failure reasons

According to the studies [8] a significant number of implementations surveyed did not deliver anywhere near the anticipated benefit or value where: 41% of companies surveyed fail to realize at least half of the business benefits they expected from their ERP systems, and 22% of implementations fail to deliver at least some measurable business benefits from their ERP solutions. In addition, (40%) surveyed companies realized major operational disruptions after

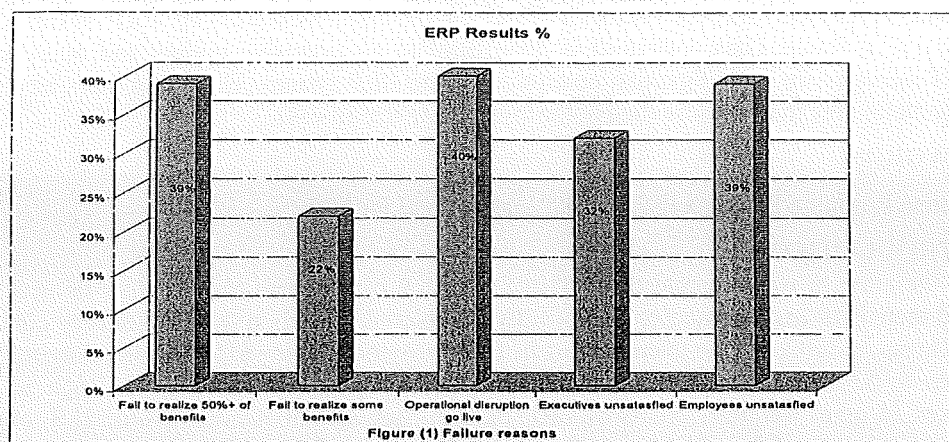
¹ Information Systems Dept. Akhbar El Yom Academy, Egypt Email: esmat99@yahoo.com, ² Information Systems Dept. Fac. of Computers & Information Cairo Univ., Egypt Email: smazen@eun.eg, ³ Information Systems Dept. Fac. of Computers & Information Cairo Univ., Egypt Email: admin@drehab.net

implementation go-live, Finally, only 68% of executives and 61% of employees are at least somewhat satisfied with their ERP solutions (Figure [1] illustrate this information).

4. The Most Obstacles of ERP Implementation

Industry statistics show that greater than 60% of ERP implementation starts historically fail. According to the many studies [8],[9],[10],[15] and our surveys on the Egyptian enterprises adopting ERP system the researchers found that there are twelve major reasons for why companies fail in implementing ERP:

- | | |
|--|---|
| 1- Lack of Top Management Commitment | 7- Misfit of Application Software with Business Processes |
| 2- Inadequate Requirements Definition | 8- Unrealistic Expectation of Benefits and Return On Investment |
| 3- Poor ERP Package Selection | 9- Inadequate Training |
| 4- Inadequate Resources | 10 - Poor Project Design and Management |
| 5- Resistance to Change/Lack of Buy-in | 11- Poor Communications |
| 6- Miscalculation of Time and Effort | 12- Poor-advised Cost Cutting |



5. Framework for Measuring ERP Implementation Success:

For identification of the Critical Success Factors (CSFs) for ERP implementation, a certain review to the related literature was performed. Some of the previous studies have also done this review before (e.g., Magnusson, J., Nilsson, A.& Carlsson, F. on Forecasting ERP Implementation Success –Towards A Grounded Framework [2]. Based on this research, a number of factors (16) were identified. This resulted in the reorganizing the factors into four categories to enhance the usability of the framework. The four categories were identified as Top Management, Project, Organization and System; and they are presented together with the underlying factors.

6. An Enhanced framework for Measuring ERP Implementation Success in the Egyptian Environment:

Based on pervious research [2], and through an extensive literary review, many other interviews with some people including IT managers having previous experience in managing ERP implementation projects and with a faculty member with a good background in this subject, the factors were then discussed and five factors were added namely:

- 1-Return on Investment (ROT) 2-Wide Commitment 3-Resistance 4-Ease to use 5-Minimal customization.

Along with a division of the known 21 factors into 4 categories (Top management, Project, Organization and System) our main concern while performing this study was to collect the factors that could be more beneficial to a developing country setting like Egypt. The final list included 21 factors figure (1), discussed later.

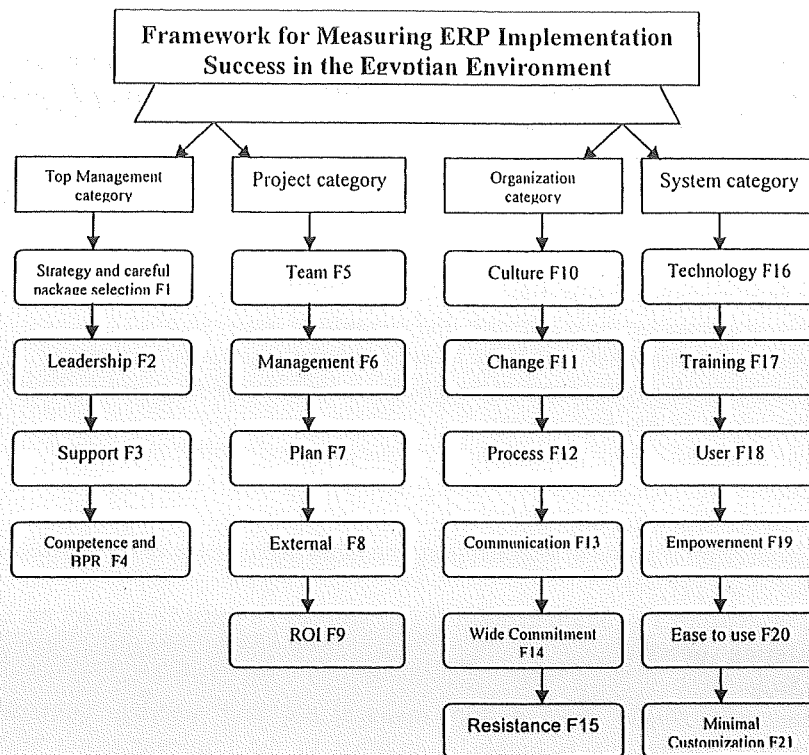


Figure (2): An Enhanced Framework for Measuring ERP Implementation Success in the Egyptian Environment

This highlights the link applying the factors and a satisfactory outcome of the project, and also for the framework to be as usable as possible; the ratio of total level of factor fulfillment to how many of the final 21 factors were fulfilled is the base of evaluation. For instance; if the responding organization fulfils 67% percent of each factor (based on 6 questions per factor, 4 questions per each factor must be replied by yes as a positive reply Except factor 15 and factor 18 (resistance and user) must be replied by no to be a positive reply) the result will be 67% That is; the probability of success of this project is 67%, and so a text describing what the organization's strengths and weaknesses will be described according to the different factors, along with a quick-list of possible future managerial actions to strengthen the identified weaknesses. A description of the 21 factors with the corresponding literary divided to pervious four categories support can be found in the following section:

6.1 Top Management Category

6.1.1 F1: Strategy and Careful package selection:

The organization should have a clear, communicated business strategy and an aligned IS/IT strategy. Most ERP software vendors make assumptions about management philosophy and business practices. Thus, buying an enterprise application means much more than purchasing software-it means buying into the software's vendor's view of best practices for many of the company's processes. For this, the company must have a detailed requirements specification before ERP package selection. [11]; [12]; [13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]; [4]; [20].

Questions that measure factor F1:

1. Does the organization have a clear, communicated business strategy?
2. Does the organization have an aligned IS/IT strategy?

3. *Does the organization buy an enterprise application?*
4. *Is there a detailed requirement specification before ERP package selection?*
5. *Does the organization have methods to evaluate employees' job performance?*
6. *Does the Organizational Strategy identify its Objectives, Initiatives and Goals?*

Factor F1: Egyptian Top Management strategy has a positive impact on ERP implementation success.

6.1.2 F2: Leadership

The organization should have a strong and committed leadership that has the ability to motivate the employees to change. [1]; [22]; [20]; [23]; [20].

Questions that measure factor F2:

- 1- *Does the organization have a strong and committed leadership?*
- 2- *Does the leadership have the ability to motivate the employees to change?*
- 3- *Does the leadership build support for goals and projects from people who do not report to him and over whom he has no authority?*
- 4- *Does the leadership have the ability to prepare a good project time plan?*
- 5- *Is the leadership capable of determining whether the candidate's style is congruent with the culture of the organization?*
- 6- *Is the leadership capable of encouraging the team members to ask questions about various portions of the project?*

Factor F2: Egyptian Top Management leadership has a positive impact on ERP implementation success.

6.1.3 F3: Support

The organization should have a top management and steering committee of the ERP Implementation project that is highly committed to the implementation and is comprised of individuals with differentiated views of the implementation. [11]; [24]; [22]; [25]; [19]; [26]; [23]; [4]; [20].

Questions that measure factor F3:

- 1- *Is there a top management committee of the ERP implementation project?*
- 2- *Is the committee committed to the implementation?*
- 3- *Is the committee comprised of individuals with differentiated views of the implementation?*
- 4- *Does the Steering Committee shift focus to emphasize adoption, utilization, measurement and refinement?*
- 5- *Does post software implementation Verify that information and knowledge are being centrally stored and shared?*
- 6- *Does post software implementation Verify that users' are actually using, benefiting by and endorsing the business software application*

Factor F3: Egyptian Top management support has a positive impact on ERP implementation success.

6.1.4 F4: Competence and BPR.

The organization should have individuals with a broad competence of ERP, BPR or other IT-related projects involved in both the steering committee and the entire project. BPR is the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service and speed" ERP systems are built on best practices that are followed in the industry. Implementing an ERP system involves reengineering the existing business processes to align the best business standards. [27]; [28]; [2].

Questions that measure factor F4:

- 1- *Are the individuals in the organization broadly competent with ERP, or other IT related projects?*
- 2- *Does BPR achieve improvements in critical, contemporary measures of performance?*
- 3- *Does BPR achieve improvements in cost, quality, service and speed?*
- 4- *Is the ERP system built on best practices that are followed in the industry?*
- 5- *Does the ERP system involve the existing processes to align the best business standards?*

6- Does the organization have the principles for business process reengineering?

Factor F4: Egyptian Competitive pressure to adopt an ERP system has a positive impact on ERP implementation success.

6.2 Project Category:

6.2.1 F5: Team.

The organization should have an Implementation project team that is comprised of individuals representing different views and perceptions of the enterprise and the enterprise system. [20];[23]; [4]; [20].

Questions that measure factor F5:

- 1- Are the individuals in the organization representing different views and perceptions of the enterprise and the enterprise system?
- 2- Does the Organization retain a subset of the installation team?
- 3- Can the organization's members enhance the ERP application, handle bottlenecks, tweak and improve the system?
- 4- Can the ERP vendor work with the implementation team to develop effective, successful customized training for the organization?
- 5- Is the support staff well qualified?
- 6- Are there criteria on which is the success of the support team measured?

Factor F5: Egyptian Project team for the ERP system has a negative impact on ERP implementation success.

6.2.2 F6: Management

The organization should have an excellent project management for the implementation project and ensure that the management does not present only a business- or technical perspective of the implementation.[14]; [24]; [22]; [25]; [26]; [23]; [4]; [20].

Questions that measure factor F6:

- 1- Does the organization have a project management for the implementation project?
- 2- Is this project management excellent?
- 3- Does the project management ensure that the management presents business and technical perspectives of the implementation?
- 4- Are there managing criteria and life cycle for a project that has a beginning and an end?
- 5- Are the project methodologies appropriate to the task and the organization?
- 6- Is there a well structured approach in complex projects to set up and approve the project?

Factor F6: Egyptian Effective project management has a positive impact on ERP implementation success.

6.2.3 F7: Plan

The organization should have a previously defined and well communicated project methodology that envelops both documentation procedures and clear performance measurements with routines for monitoring progress [26].

Questions that measure factor F7:

1. Does the organization have previously defined and well communicated project methodology?
2. Does the project methodology envelop documentation procedures?
3. Does the project methodology envelop clear performance measurements with routines for monitoring progress?
4. Does the organization determine which Partners or Collaborators will be needed?
5. Does the organization determine if any Specific Information or Advice will be needed?
6. Are there guidelines to define each one's duty?

Factor F7: Egyptian Project planning to best of practices of an ERP system has a positive impact on ERP implementation success.

6.2.4 F8: External

The organization should have an ability to manage the influence of external consultants in the implementation project and also be able to optimally transfer the knowledge from the consultants into the organization [23]; [20].

Questions that measure factor F8:

- 1- Is the organization capable of managing the influences of external consultants in the implementation project?
- 2- Is it capable of optimally transferring the knowledge from the consultants into the organization?
- 3- Do Project management consultants distance themselves from politics, conflicts with other managers and chit-chats with company employees?
- 4- Are the external project managers highly efficient?
- 5- Does the external project manager work with the team members more than 6 months?
- 6- Are internal project managers better than external ones?

Factor F8: Egyptian External support and consultants has a positive impact on ERP implementation success.

6.2.5 F9: Return On Investment (ROI)

The organization should have an Achieving expected payoff (Return On Investment) Calculating return on Investment, (ROT) , which is based on comparing the costs of implementation to the benefits that you will receive by the integration of an ERP software package into your business structure. While the direct costs of the software itself, installation, training, and support can generally be quantified by the vendor, there are other indirect costs that will be generated within the business, such as reduced productivity during the integration process and training sessions. Langenwaller claims that the percentage of ERP implementations that can be classified as "failures" ranges from 40% to 60% or higher [29]. Failure defined as an implementation that does not achieve the ROI identified in the project approval phase and finds that failure rates are in the range of 60–90%.

Questions that measure factor F9:

1. Does the organization have an expected payoff (return on investment)?
2. Are the indirect costs of the software (as the reduced productivity during the integration process and training sessions) generated within the business?
3. Isn't ROI too complicated for most learning and HR professionals?
4. Is it true that the ROI process does not reveal program weaknesses or strengths?
5. Is it possible to isolate the effects of any program from other factors?
6. Does the ROI process result in recommendations for improvement?

F9: Egyptian organizational Return On Investment (ROT) has a negative impact on ERP implementation success.

6.3 Organization category:

6.3.1 F10: Culture

The organization should have a business culture that highlights the importance of learning, knowledge, past experience and change, as well as a strategy for knowledge management. The following points are identified to be associated with ERP implementation success [30]; [31]:1- Reliance on information, 2- Design of the general structure,3-. Design of Lateral links, 4- Centralization of decisions. 5- High Context communication: Egyptian employees have highly collective nature. This predisposes culture against computer based communications because these media mutate the group effect. Workers in collectivistic nature favor high personal meetings and phone calls on IT communication technologies. [13]; [14] ;[17],[18]; [20]; [32];[33]; [20];[30];[31][1];[5].

Questions that measure factor F10:

1. Does the organization have a business culture highlighting the importance of learning, knowledge, past experience and change?
2. Does the organization have a strategy for knowledge management?
3. Is the information perceived to be an organizational resource in the Egyptian organizations culture?
4. Do top managers rely on the implemented information system?
5. Is cooperation across different functional areas in the Egyptian organization entailed by ERP achieved?
6. Does the Egyptian organization have an effective design of lateral links?

Factor F10: Egyptian organizational culture has a negative impact on ERP implementation success.

6.3.2 F11: Change.

The organization should have a fundamental willingness and readiness for change as well as an explicit change management strategy. [11]; [12]; [1];[24];[22]; [25]; [20]; [23] ;[4] ;[20].

Questions that measure factor F11:

1. *Does the organization have willingness and readiness for change?*
2. *Does the organization have an explicit change management strategy?*
3. *Does the organization have common pitfalls when implementing organizational change?*
4. *Does the organization have tools to minimize resistance from middle management in implementing this change?*
5. *Does the organization have buzzword should to implement?*
6. *Does the organization get benefit from a consultant?*

Factor F11: Egyptian organizational change has a negative impact on ERP implementation success.

6.3.3 F12: Process

The organization should have a high level of process-maturity and explicit guidelines for process management. [1]; [13]; [28];[14]; [22]; [23].

Questions that measure factor F12:

- 1- *Are there guidelines for process management within the organization?*
- 2- *Are there criteria for designing and innovating the overall work system?*
- 3- *Are there criteria by which the work systems and key work processes relate to and capitalize on your core competencies?*
- 4- *Are there criteria to ensure that the day-to-day operation of these processes meets their key process requirements?*
- 5- *Are there measures to control the overall costs of your work processes and prevent defects, service errors, and rework?*
- 6- *Are there measures taken so that work process improvements and lessons learned shared with other units and processes to drive learning and innovation?*

Factor F12: Egyptian organizational process has a positive impact on ERP implementation success.

6.3.4 F13: Communication.

The organization should have a detailed communication plan and strategy that ensures the successful communication of project plan and progress to all relevant stakeholders[11]; [12];[1];[22]; [19]; [20]; [23]; [20].

Questions that measure factor F13:

1. *Is there a detailed communication plan and strategy within the organization?*
2. *Does this plan (if present) ensure the successful communication of project plan and progress to all relevant stakeholders?*
3. *Is there Critical Strategic Communications Practices within the organization?*
4. *Are there criteria for assessing Current Performance and Capacity?*
5. *Are there criteria to Identify Possible Levels of Practice?*
6. *Are there criteria to Identify Areas for Improvement?*

Factor F13: Egyptian organizational communication has a positive impact on ERP implementation success.

6.3.5 F14: Wide Commitment.

The organization should have a ERP systems are enterprise-wide IS that integrate information and information based processes within and across all functional areas in an organization, it's imperative to get support from all functional segments of the organization (Every person and department is responsible accountable for the overall system and key users from different departments are ensured to commit to the project implementation without being called back to their prior functional job position frequently [1].

Questions that measure factor F14:

1. Does the organization have an ERP system that integrates information and information based process within and across all functional areas in the organization?
2. Are the persons and departments responsible for the overall system?
3. Are the key users from different departments ensured to commit to the project implementation?
4. Are these departments frequently being called back to their prior functional job position?
5. Is there a prepared list of what is required from each of the functional areas of the company?
6. Is the global team well trained in the organization?

F14: Egyptian Company wide support has a positive impact on ERP implementation success.

6.3.6 F15: Resistance.

The organization should have an ERP systems usually introduce large-scale changes that can cause resistance, which may in turn decrease the expected benefits of the system. Previous IS research has made substantial progress in understanding how resistance affect IS success [34].

Questions that measure factor F15:

1. Does the ERP system introduce large-scale changes that cause resistance?
2. Does this resistance decrease the expected benefits of the system?
3. Is there in the company any parochial self interest?
4. Is there in the company any misunderstanding and lack of trust?
5. Is there in the company low tolerance for change?
6. Does the company introduce a single method for assessment?

Factor F15: Egyptian Organizational resistance has a negative impact with ERP implementation success.

6.4 System Category:

6.4.1 F16: Technology.

The organization should have a clear understanding of the existing legacy environment and the technological aspects involved in the implementation of the ERP system.[1];[13];[28];[17],[18]; [22]; [25]; [20]; [4].

Questions that measure factor F16:

1. Is the organization technology integrated with the ERP solution to boost the organization's return on investment?
2. Is there a clear understanding of the existing legacy environment?
3. Is there are a clear technological aspects involved in the implementation of the ERP system?
4. Is there a well performed Hardware and networking for ERP system?
5. Does the ERP software vendor generally certify which hardware (and hardware configurations) must be used to run the ERP system?
6. Is the Hardware selection driven by the firm's choice of an ERP software package?

Factor F16: Egyptian Organizational Technology has a positive impact with ERP Implementation success.

6.4.2 F17: Training.

The organization should have a clear educational strategy concerning the ERP implementation that involves routines for early hands on training for the employees. [11];[1]; [22]; [23]; [4]; [20].

Questions that measure factor F17:

1. Is there a clear educational strategy concerning the ERP implementation?
2. Does this strategy involve routines for early hands on training for the employees?
3. Are there definite training issues in implementing a new ERP system?
4. Is the training curriculum designed for the new organizational roles and tasks?
5. Is the training customized for your specific business?
6. Are the learning materials developed by experienced instructional designers and tailored to different learning styles

Factor F17: Egyptian training has a positive impact on ERP implementation success.

6.4.3 F18: User

The organization should have an implementation process that strives for a high level of user acceptance early on through the use of constant presumptive end-user consultations. [22]; [19]; [26]; [23]; [20].

Questions that measure factor F18:

1. *Does the implementation process not strive for a high level of user acceptance?*
2. *Does the organization not use constant presumptive end-user consultations?*
3. *Are there differences in user acceptance of an ERP system between users belonging to different departments?*
4. *Are there differences in user acceptance of an ERP system between users in different positions?*
5. *Are there differences in user acceptance of an ERP system between user tenure?*
6. *Are there differences in user acceptance of an ERP system between user ERP implementation experiences?*

Factor F18: Egyptian Users' involvement has a positive impact on ERP implementation success.

6.4.4 F19: Empowerment

The organization should have a high level of implementation process transparency and a staff policy that empowers team members, end-users and management [11]; [25];[20].

Questions that measure factor F19:

1. *Does the organization have a high level of implementation process transparency?*
2. *Does the organization have a staff policy that empowers team members, end-users and management?*
3. *Is the user interface consistent throughout the application?*
4. *Is the software licensed?*
5. *Can a non-programmer develop a new business process in the system?*
6. *Are there criteria to know that the proposed solution support the style of manufacturing?*

Factor F19: Egyptian Empowerment has a positive impact on ERP implementation success.

6.4.5 F20: Ease to use.

The organization should have a strategy to select the ERP system to be ease of use. Ease of to use is the degree to which a particular system is perceived to be relatively free from physical and mental effort. Previous is Research found that ease of use has an impact on the intensions to use ERP systems [35].

Questions that measure factor F20:

1. *Does the organization have a strategy to select the ERP system to be easy to use?*
2. *Do the User concerns and issues set the balance and priorities in creating Usability goals?*
3. *Are the ERP vendors continually improving their products with new features?*
4. *Does the organization rank highest in decision criteria especially when one product looks a lot easier to use than another.*
5. *Does the application content of an ERP product exceed the expectations of buyers?*
6. *Does the organization select the ERP system that is easy to train new employees and easier to support?*

Factor F20: Egyptian Perceptions on the ease of use of the ERP system has a positive impact with ERP implementation success.

6.4.6 F21: Minimal customization.

Is using the vendors' code as much as possible. Taming the package causes extra additional customization costs, inability to benefit from vendor software maintenance and upgrades, and inability to benefit from the standard best practices encapsulated in the package [1].

Questions that measure factor F21:

1. *Does the organization minimize customazitation of the ERP system?*
2. *Does the organization differentiate between the two key types of customizations - usability and functionality?*
3. *Does the organization understand that the issue of customization has been plaguing the ERP software industry since its inception?*

4. Do the Organizations deploy an ERP that is being customized to fit their needs only to then discover that the cost of applying maintenance?
5. Does the organization know the Reasons for ERP system customization prior to "going-live" and after "going-live"?
6. Do the organization Factors include the need to address strategic and cost implications, organizational adaptation to ERP functionality, deploying strategies to minimize customizations.

Factor F21: Egyptian system Minimal customization of the ERP system has a positive impact with ERP implementation success.

During the industrial feedback sessions, the first draft of the conceptual framework, and the need for the user to see some sort of structure in the 21 factors all were identified (see Figure 2). This resulted in the reorganizing the factors into four categories that would enhance and facilitate the way to use this framework. The reliability and validity of such categorization was then considered to be irrelevant because there was a need for the future user to a more sharp focus. These four categories were identified as Top Management, Project, Organization and System; and they are presented together with the underlying factors in Figure 1 below. As previously described the fulfillment of the factors is in the basic outline of the framework measured by 6 questions each, resulting in a total of 126 questions. These questions are described over here and they constantly revised and tested.

Based on the above arguments and the suggested research framework we can summarize the following factors are postulated:

Factors F1, F2, F3, F4, F5, F6, F7, F8, F12, F13, F14, F16, F17, F18, F19, F20, F21 has a positive impact with ERP implementation success. But factors F9, F10, F11 and F15 have a negative impact with ERP implementation success in Egyptian environment.

7. RESULTS:

After Applying our framework about the Critical Success Factors (CSFs) that ensure ERP systems implementation success on three big organizations in Egypt we got the following results:

O1 : achieved 96% success rate .(Success ERP project) figure (3)

O2 : achieved 45% success rate (Failed ERP project) figure (4)

O3: achieved 93% success rate (Success ERP project) figure (5)

7.1 Results of O1:

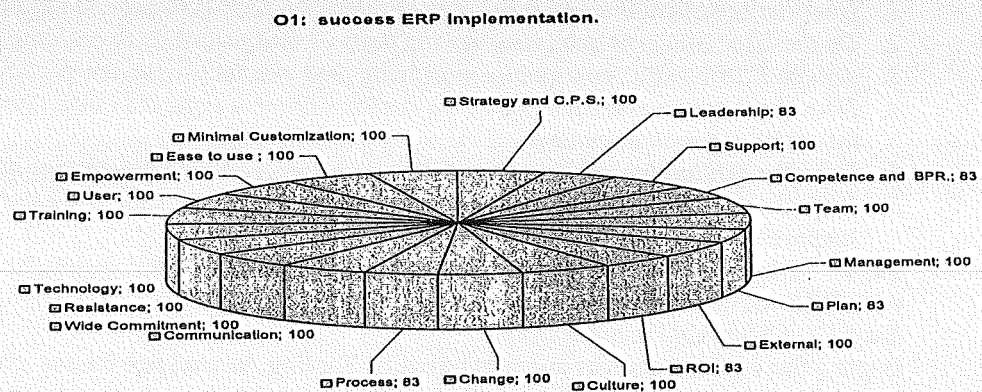


Figure (3)

From figure (3) we conclude that most factors (16 factors out of 21) 76% achieved 100% success rate and the remaining factors 23% achieved 83% success rate indicating the success ERP implementation in O1 Organization as the following:

1. For the top management category we find that the two factors achieve 100% success rate especially the Strategy and careful package selection factor (F1) and the support factor (F2) that refer to the good management in this organization (O1).
2. For the project category we find that three factors achieve 100% success rate ; team factor (F5) ,the management factor (F6) and the external factor (F8) that also refer to the success the project category in this organization.
3. For the organization category we find that all related factors (Culture (F10) ,Change (F11) ,Communication (F13),Wide commitment (F14) and Resistance factor (F15)) except only one factor (Process factor (F12)) achieve the 100% success rate that refers to the high success rate at the organization category in this organization.
4. For the system category we find that all related factors (Technology (F16), Training (F17),User (F18), Empowerment (F19) ,Ease to use (F20) and Minimal Customization (F21)) all achieve 100% success rate that refer to the complete success rate for the system category at this organization .

7.2 Results of O2:

From figure (4) we conclude that:

- 29% of the factors achieved 67% success rate (6 factors out of 21),
- 19% of the factors achieved 50% success rate (4 factors out of 21),
- 19% of the factors achieved 33% success rate (4 factors out of 21),
- 10% of the factors achieved 17% success rate (2 factors out of 21),
- 5% of the factors achieved 100% success rate (one factor out of 21),
- 5% of the factors achieved 83% success rate (one factors out of 21),
- 14% of the factors achieved 0% success rate (3 factors out of 21),

O2: failed ERP Implementation.

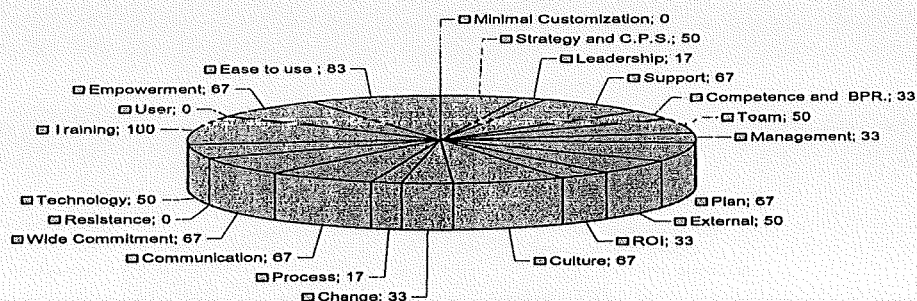


Figure (4)

Indicating failure ERP implementation in O2 Organization as the following:

1. For the top management category we find that only the support factor (F3) achieve exactly the percentage for the successes any factor (67% success rate) but the other factors (Strategy and careful package selection (F1), Leadership (F2), Competence and BPR. (F4)) Do not achieve the percentage for the success of this category at this organization (O2).
2. For the project category that include five factors we find that only one factor achieve exactly the percentage for the successes (67% success rate) plan factor (F7) but the other factors: team factor (F5), the management factor (F6), the external factor (F8) and the ROI factor (F9) do not achieve the success rate that perform the success of this category in this organization (O2).
3. For the organization category we find that only 50% of related factors (Culture (F10), Communication (F13), Wide Commitment (F14) achieve exactly the percentage for the successes (67% success rate) but the other factors Change (F11) ,process (F12) , resistance

(F15) do not achieve the success rate that perform to the success of this category in this organization (O2).

4. For the system category that include 6 factors we find that only three factors achieve success rate as the following: (Training (F17) achieve 100%, Empowerment (F19) achieve exactly the percentage for the successes (67% success rate), Ease to use (F20) achieve 83% success rate but the other factors (Technology (F16), User (F18) , Minimal Customization (F21) do not achieve the success rate that perform to the success of this category in this organization (O2).

7.3 Results of O3:

From figure (5) we conclude that most factors (16 factor out of 21) 76% achieved 100% success rate , two factors (2 factor out of 21) 14% achieved 83% success rate and the remaining factors (3 out 21) 5% achieved 67% success rate indicating the success ERP implementation in O3 organization.

1. For the top management category we find that all factors achieve 100% success rate that refer to the very good management in this organization (O3).
2. For the project category that include five factors we find that two factors achieve 67% success rate ; team factor (F5) ,the external factor(F8) and two factors achieve 83% success rate (ROI (F9), Minimal customization (F6) that also refer to the success the project category in this organization(O3).
3. For the organization category we find that all related factors achieve 100% success rate except the Change (F11) factor that achieve 83% that refers to the high success rate at the organization category in this organization .
4. For the system category we find that all related factors achieve 100% success rate except the Minimal customization (F21) that achieves 83% success rate and refer to the complete success rate for the system category at this organization.

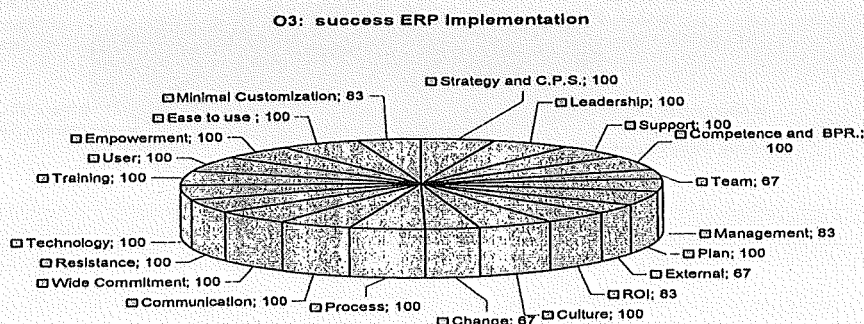


Figure (5)

Figure (6) shows the summary of all the pervious results according to our framework for Critical Success Factors (CSFs) for ERP implementation. this figure shows that the known 21 factors are divided into 4 categories (Top management, Project, Organization and System) that first organization O1 achieve 96% as average success rate, second organization O2 achieve 45%, and last organization achieve 83% average success rate.

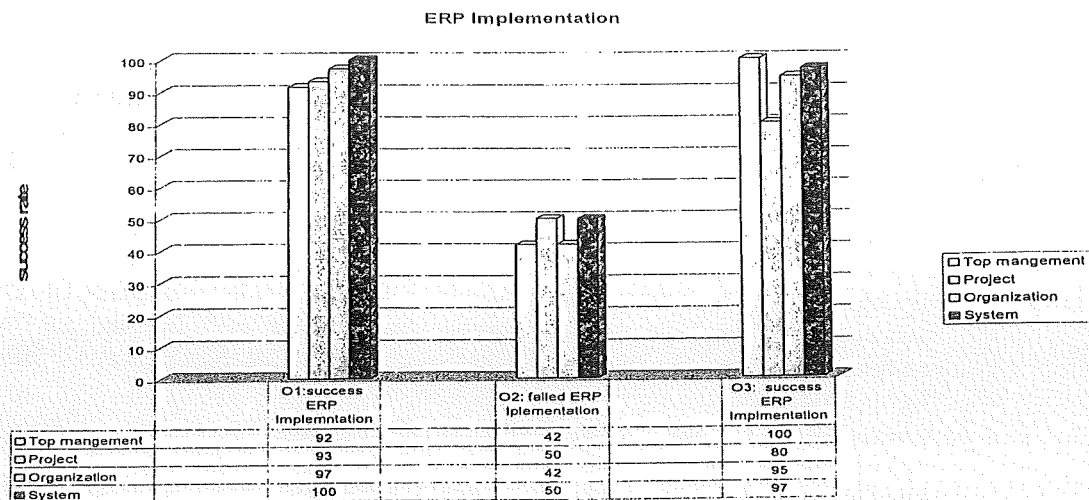


Figure (6)

8. Conclusion

Egypt's expenditure on ERP and other enterprise systems is growing, and these systems can undoubtedly deliver benefits to organizations in developing countries [36]. However, high failure rates continue to block the delivery of such benefits. Research to date, though, often appears partial, focusing on only some aspects of system outcome and/or focusing only on certain specific implementation factors.

Through an extensive literary review and many other interviews with IT managers having previous experience in managing ERP implementation Five implementation factors were added in this research (1- Return on Investment (ROT) 2-Wide Commitment 3-Resistance 4-Ease to use 5- Minimal customization).

This research concentrated on practical testing of the framework. This is mainly done through testing the reliability and validity of the framework by using it with corresponding 126 questions.

These questions were distributed as 6 question per 21 factors grouped into 4 categories (Top management, Project, Organization and System) as a model, After that we used the model for following a smaller number of organizations through an implementation. After this we got statistical reports and graphs about the outcome from each factor corresponding to the organizations under study. And we got statistical reports and graphs about the outcome of the 4 categories (Top management, Project, Organization and System) for all organizations under study indicating the average success rate for ERP Implementation of each organization.

This paper research adds to the growing body of knowledge on ERP implementation, helping studying ERP implementation in the Egyptian settings. For practitioners, it provides ERP adopters with a self-evaluation tool that helps them to monitor and predict the likelihood of project success.

9. FUTURE WORK

In terms of future research, due to the limited number of existing models for ERP implementation, more research is still needed to investigate the correlation between the Critical Success Factors (CSFs) for ERP implementation and the failure or success of the ERP implementation. Also, the generalisability of the framework described in this paper to other industries, larger organizations, and technology driven ERP investments should be investigated.

Also our future study will be concentrated on building an expert system for the suggested framework to represent a self-assessment tool for ERP implementers in Egypt. Combining both detailed case study and a large survey would be an ideal method for researchers in the ERP field.

10. ACKNOWLEDGEMENT

We would like to thank all the contributors for the success of this paper.

11. REFERENCES

1. M.H. Rasmy, A.Tharwat and S. Ashraf "Enterprise Resource Planning (ERP) Implementation in the Egyptian Organizational", Cairo University, Egypt, 2005.
2. Magnusson, J., Nilsson, A.& Carlsson, F.,"Forecasting Erp Implementation Success – Towards A Grounded Framework" University Of Gothenburg, Department Of Informatics.
3. T.L.Griffith,L. Zammuto,and L.Aiman-Smith, "Why new technologies fail?" Industrial management:29-34,1999.
4. Umble, E. J. (2003) Avoiding ERP Implementation failure, Industrial Management February.
5. Abdelmoniem, E.M, Mazen,A.S. Hassanein, E.E. (2011)."Management & Governance of Successful ERP Implementation in Egyptian Environment" Fifth International Conference on Intelligent Computing and Information Systems (of ACM, ICICIS,3 JULY 2011, Cairo, Egypt,231:250.
6. Abdelghaffar, H. and Abdel Azim, R. H. "Significant factors influencing ERP implementation in large organizations", Evidence from Egypt European, Mediterranean & Middle Eastern Conference on Information Systems 2010 April 12-13 2010, Abu Dhabi, UAE
7. Smith, A. (1982) The Wealth of Nations: Books I-III, Harmondsworth: Penguin.Standish Group International Inc (1999) Chaos: A Recipe for Success, The Standish Group International Inc.
8. S.Alter, "A general, yet useful theory of information systems, Communications of the Association for Information Systems" vol.1, no.13,pp. 37, 1999.
9. (2010) ERP failure: New research and statistics: E-Business Insight - ERP, CRM and Supply Chain Management, 2005-03-12, http://www.sysoptima.com/erp/cost_of_erp.Php.
10. Richard G. Ligus is President of Rockford Consulting Group located in Rockford,IL., <http://www.articlesbase.com/management-articles/the-12-cardinal-sins-of-erp-implementation> Edit Article | Posted: Feb 01, 2009.
11. Aladwani, A.M. 2001. Change management strategies for successful ERP implementation. Business Process management journal, 7(3):266-275.
12. Al-Mashari, M. & Zairi, M. 2000. Information and business process equality: the case of SAP R/3 implementation. Electronic journal on information systems in developing countries, 2.
13. Al-Mashari, M. 2001. Process orientation through Enterprise Resource Planning (ERP): A review of critical issues. Knowledge and Process Management, 8(3):175-185.
14. Cooke-Davies, T. 2002. The "real" success factors on projects. International journal of project management, 20:185-190.
15. Davenport, T. H. (2000) Mission critical: Realizing the promise of Enterprise Systems, Harvard Business School Press.
16. Donovan, M. 1999. Successful ERP implementation the first time. Midrange ERP August.
17. Holland, C. & Light, B. (1999). A critical success factors model for ERP implementation. IEEE Software May/June 30-35.
18. Holland, C. and Light, B. (1999) A framework for understanding success and failure in Enterprise Resource Planning System Implementation. 7th European Conference on Information Systems, Copenhagen.
19. Pinto, J.K. & Slevin, D.P. 1987. Balancing strategy and tactics in project implementation. Sloan Management review:33-41.
20. Sarker, S. & Lee, A.S. 2003. Using a case study to test the role of three key social enablers in ERP implementation, Information and management. In press Schneider, P. 1999. Wanted:.
21. Whyte, D. & Fortune, J. 2002. Current practice in project management – an empirical study.

22. Mandal, P. & Gunasekaran, A. 2003. Issues in implementing ERP: A case study. *European Journal of operational Research* 146:274-283.
23. Skog, W. & Legge, M. 2002. Evaluating Enterprise Resource Planning (ERP) Systems using an interpretive approach. *Knowledge and Process Management*, 9(2):72-82.
24. Wokingham Kerzner, H. 1987. in search of excellence in project management. *Journal of systems management*: 30-39 software technology, 44:53-62.
25. Parr, A. & Shanks, G. 2000. A model for ERP project implementation. *Journal of Information Technology*, 15:289-303.
26. Procaccino, J.D., Verner, J.M., Overmeyer, S.P. & Darter, M.E. 2002. Case study: factors for early prediction of software development success.
27. Hammer, M. and J. Champy. 2001, 'Reengineering the Corporation: A Manifesto for Business Revolution', Harper Business, New York, NY, USA, work organization? *Strategic Change*, 11:263-270.
28. Bingi, P., Sharma, M., & Godla, J.K. (1999), "Critical issues affecting an ERP implementation", *Information Systems Management*, (Summer), 7-14.
29. G. Langenwalter, *Enterprise Resources Planning and Beyond: Integrating Your Entire Organization*, St. Lucie Press, Boca Raton, FL, 2000.
30. Burn, J., Davison, R. and Jordan, E. (1997), "The Information Society – A Cultural Fallacy," *Failures and lessons learned in information technology management*, vol. 1, 219-232.
31. Zhang, Liang, Matthew, K. O., Lee, Zhe Zhang and Banerjee, Probir. (2003), "Critical Success Factors of Enterprise Resource Planning Systems Implementation Success in China", in *Proceedings of the 36th Hawaii International Conference on System Sciences*.
32. Soffer, P., Golany, B. & Dori, D. 2003. ERP Modeling: a comprehensive approach. *Information systems* 28:673-690.
33. Sumner, M. 1999. Critical Success factors in enterprise wide information management systems projects. *Proceedings of the ACM-SIGCPR Conference*:297-303.
34. Cooper, R., & Zmud, R. (1990). Information technology implementation research: a technological diffusion approach. *Management Science*, 36(2), 123-139.
35. Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.
36. A.Molla, and A.Bhalla, "Business transformation through ERP: a case study of an Asian company", *Journal of Information Technology Case and Application Research*, vol. 8, no. 1, pp. 34-54, 2006.

Handwritten text at the top of the page, possibly a title or header, which is mostly illegible due to fading and bleed-through.

Main body of handwritten text, consisting of several paragraphs. The text is extremely faded and mostly illegible, appearing as light grey shapes against the white background.

Handwritten text at the bottom of the page, likely a footer or concluding remarks, also mostly illegible.

The Role of the Pre-development Activities in Information Systems Projects' Failure

Mohamed A. Hamada⁽¹⁾

Sherif A. Mazen⁽²⁾

Ehab E. Hassanein⁽³⁾

Abstract

Information systems (IS) development projects have high failure rates, lots of researches discussed IS projects failures and listed a huge number of failure reasons, despite that IS projects still have considerably high failure rates. This research looks for the sources of the failures as a new approach to decrease the failure probability of IS projects. Through investigating the failure reasons of IS projects, it has been detected that, most of failure reasons are attributed to the pre-development activities, which begin with the System Service Request (SSR) to initiate the projects and end with developing the project plan to represent a road map to explain how the project will progress through all its phases. Pre-development activities are focal points which have ability to improve IS projects success. So, an avoidance strategy includes many critical success factors for each pre-development activity is introduced as a swift solution for the research problem.

Keywords: Failures of IS projects, Pre-development activities, Project planning, Avoidance Strategy,

1. Introduction:

Information systems development is a complex process and comprises many challenges and difficulties in all development activities with high risks of such projects end in failure [22]. There is a significant body of evidence that many information systems (IS) implementation projects end in failure, the failure rate for major systems appears to linger around 70% [28]. Projects failures can take many forms and have consistently been defined in terms of projects that could not be completed on time or within budget or that did not meet stakeholders' expectations or that were cancelled [39].

High failure rate of IS projects is considered a major problem because the occurrence of enormous costs and losses [40]. Recent studies of IS project failure reveal a number of interesting themes, for instance, failure could be caused by multiple factors, such as unrealistic expectations, lack of resources, uncooperative customers, and weak management of contractors. Failure could also be caused by the political rivalry in organizations; other themes also encompass the technical aspects of the system, such as failure in meeting predefined design objectives. Nevertheless, many IS studies indicate that failure is largely due to organizational and social, rather than technical factors [28]. Lots of researches investigated the failure of IS project and listed a huge number of projects failure reasons, but the IS projects still fail and have high failure rate. This problem creates an obligation case to search for the origins of the failure as a new approach to decrease the failure probability of IS projects and improve the percentage of IS projects success.

¹ PhD student, Information Systems Department, Faculty of Computers and Information, Cairo University

² Associated professor of Information Systems, Information Systems Department, Faculty of Computers and Information, Cairo University

³ Assistant professor of Information Systems, Information Systems Department, Faculty of Computers and Information, Cairo University

The objective of this paper is to clarify the role of the pre-development activities in the failure of IS projects through analyzing the recent statistics and researches that demonstrate the failure reasons of IS projects, also shed light on the pre-development activities and setting an avoidance strategy for each pre-development activity as a swift solution to decrease the failure of IS projects. This paper is organized as follows: section two exposes the main failure reasons which have a relation with the pre-development activities, section three describes the research methodology, Section four investigates the pre-development activities and illustrates the correlation between them and the projects failure reasons, Section five introduces the analysis of FBI case study to explain the importance of the pre-development activities in projects success achievement, Section six explains the main factors of the avoidance strategy for each pre-development activity as solution to decrease IS Projects failure, Section seven conclusions with suggestions for the future research.

2. Why Information Systems Projects Fail?

In general Information systems projects have a high failure rate, completing projects on time, within budget, and with quality is a major challenge [20]. Standish Group survey proved that 65% of IS projects which started in 2006 were failed [5]. Another study conducted in the UK by Oxford University and Computer Weekly in 2003 reported, strikingly, that only 16% of the IS projects reviewed were considered successful. Projects failure reasons have been attributed to technological difficulties, organizational and functional problems, managerial issues, and many other reasons. IS project failure is defined as any project that was set to support the operations of an organization by exploiting the resources of information technology that fails to deliver the intended output within the originally allocated cost, time schedule, and accepted level of quality, as well as fails to satisfying the stakeholders and being accepted by the end users [1].

Normally, IS projects are often difficult to estimate and manage, some projects are canceled or reduced in scope because of overruns in cost and time or failure to produce anticipated benefits [31]. It's widely recognized that poor project planning plays a major role as one of the significant causes of project failure [14]. The fundamental focus of project management has been to deliver projects on time, on budget and meet specifications with all required functions and features initially specified. However, many major projects still fail to meet these targets, especially on cost and schedule. The reasons for failure have often been blamed on poor project definition, incomplete information, poor productivity, inadequate communications, uncertainties around labor and material costs, and the failure to use timely and appropriate project management practices and controls [36]. So, IS projects are considered success if they achieved the following issues:-

- 1- If they accomplish its business objectives (strategic, technical and operational).
- 2- If they are satisfy and profitable for the sponsor/owner and contractors.
- 3- If they meet its expectation requirements and objectives with a reasonable quality.
- 4- If they are produced to specification, within budget and on time.

Lots of researches listed a large number of failure reasons for IS projects [14], [15], [34], [1], [5], [19], but this research initiates an effort to look for the origins of the IS projects failure. It's important recognize that, there is no project suffered from just one failure reason, but all projects had multiple troubles and the failure reasons are related to each other. Overlaps between the projects failure reasons are evidential through examining the following statistics in (Table1) that published in (2008) by the two famous independent research organizations for IS/IT projects [29], and (Table 2) wherever Cerpa and Verner in (2009) demonstrated the failure reasons of seventy failed projects developed in house at U.S., Australia, and Chile[5]. Symbol column is added for both two tables to indicate the failure reasons at the residual sections of this research.

Table: 1, Failure Reasons Statistics for IS Projects (Source: PM Hut, 2008)

Symbol	The Standish Group Statistics	Percentage
A1	Lack of User Involvement	12.4%
A2	Lack of Resources	10.6%
A3	Unrealistic Expectations	9.9%
A4	Lack of Executive Support	9.3%
A5	Changing Requirements & Specifications	8.7%
A6	Lack of Planning	8.1%
A7	Didn't Need It Any Longer	7.5%
A8	Lack of IT Management	6.2%
A9	Technology Illiteracy	4.3%
A10	Incomplete Requirements	13.1%
A11	Other	9.9%
Symbol	Bull Survey Statistics	Percentage
B1	Bad Communications	57%
B2	Lack of Planning	39%
B3	Poor Quality Control	35%
B4	Missing Interim Deliverables	34%
B5	Poor Budget Management	29%
B6	Poor Project Management	20%

Table: 2, Failure reasons for seventy failed projects, (Source: Cerpa and Verner, 2009)

Symbol	Failure reasons	Percentage
C1	Delivery date impacted the development process	93.9%
C2	Project under-estimated	83.7%
C3	Risks were not re-assessed, controlled, or managed through the project	73.4%
C4	Staff were not rewarded for working long hours	81.6%
C5	Delivery decision made without adequate requirements information	83.7%
C6	Staff had an unpleasant experience working on the project	83.7%
C7	Customers/	69.4%
C8	Risk not incorporated into the project plan	65.3%
C9	Change control not monitored, nor dealt with effectively	63.3%
C10	Customer/Users had unrealistic expectations	69.4%
C11	Process did not have reviews at the end of each phase	75.5%
C12	Development Methodology was inappropriate for the project	71.4%
C13	Aggressive schedule affected team motivation	69.4%
C14	Scope changed during the project	67.3%
C15	Schedule had a negative effect on team member's life	71.4%
C16	Project had inadequate staff to meet the schedule	63.3%
C17	Staff added late to meet an aggressive schedule	61.2%
C18	Users did not make adequate time available for requirements gathering	61.2%

3. Research Methodology

A quantitative research methodology was followed to seek illumination and understanding through extrapolation of the findings at hand, this methodology was selected to gain an in-depth understanding. The research methodology based on extrapolation the literature review supported with a quantitative survey for the previous researches that investigated the failure and success of information systems projects. Therefore, statistics of the failure reasons for IS projects are examined in order to explore the origins of the failures for IS projects, this with shed light and focus on the pre-development activities as a critical success factors for IS projects which begin with the System Service Request (SSR) to initiate the projects and end with the developing the Project Plan. The research methodology is used to explain and prove many issues in this research; the first issue, there is no project suffered from just one failure reason. The second issue, most of all IS project failure reasons are originated from the pre-development activities. The third issue, it is very important to highlight and focus on the pre-development activities to decrease the failure probability for IS projects. Fourth, an avoidance strategy can be taken as a swift solution to improve the IS project success.

4. Pre-development Activities for IS projects:

Pre-development activities are derived from the initiation and planning phases for IS projects development. They are the main critical success factors for IS projects, If these activities are implemented accurately, the projects will achieve success in high percentages [17]. They begin with the System Service Request (SSR) to initiate the projects and end with developing the project plan which represents a road map to explain how the project will progress through all project phases and outlined objectives. High risk of the pre-development activities comes from the certainty of projects fail if these activities missed the accuracy and are performed in improper form, furthermore IS project will lose the great benefits from the optimal pre-project planning which yielding the following features; Fewer scope changes, Increased predictability of cost and schedule, Improved operational performance, Better achievement of business goals, and Better definition of risks [12].

In the fact, the previous researches in the area of IS projects didn't collect the pre-development activities as an integrated framework but they discussed it in many scattered areas [2], [24], [35], [16], [39], [32], [7], [11], [38]. In this research, the pre-development activities will be grouped and classified to prove the great relation between the failures of IS projects and the way in which these activities are implemented. In the context of surveying the pre-development activities which play an important role to determine the fate of the IS projects success or failure, it concluded with extracting Fifteen critical pre-development activities as the following:-

4.1-System Service Request (SSR)

Most IS projects are initiated in response to a request from a client or manager who has an idea or need [24], System Service Request (SSR) is a standard form for requesting development or maintenance of an information system within an organization, it includes the contact person, a problem statement, a service request statement, and the relational contact information [16].

4.2-Project Scope Definition

Project scope definition is the process by which projects are defined and prepared for execution; Poor specification can lead to numerous requirement changes and scope creep [40]. Unclear and incomplete objectives have been shown to contribute to the perception of project failure [3]. Project Management Institute (PMP) listed the main features to define the project scope; Project justification, Project product, Project deliverables, and Project objectives [30].

4.3- Project Feasibility

Feasibility study determines whether the requested information system makes economic and operational sense for an organization or not, it considers a crucial matter to determine the project proceeding which includes five feasibility tests; Economic feasibility, Technical feasibility, Operational feasibility, Schedule feasibility, Legal and Political feasibility [16], [38], basic data for the feasibility study is obtained from the client through a series of queries, questions, and meetings, wherein the client provides some of the research, other data and facts need to be gained from a variety of sources.

4.4- Commercial Off-The-Shelf (COTS)

COTS have the ability to reduce time and cost of IS project development, they enable software buyers to acquire software made up of components which have been tested many times by other users; hence ensuring improved software quality. In order to realize the benefits which COTS bring to software development, it is imperative that the "right" COTS is selected for a project, because selecting inappropriate COTS may result in increased time and cost of software development. A COTS product can consist of one or many components so, the selection process of COTS is a complex decision-making [37]. It can cause many problems for IS project through the following reasons that must be avoided as a failure reasons;

1. Uncertainty: There is uncertainty about the requirements of the software under development, about the correctness of the available information, and about the impact of the COTS on the quality of the system.
2. Complexity of the decision-making problem: COTS selection decisions are multileveled, different decisions are required to be made at different stages of the COTS selection process
3. The use of commercial-off-the-shelf (COTS) products creates a software integration problem: whether a single COTS software component is being integrated into a software system
4. Multiple Stakeholders: The COTS selection process usually involves multiple stakeholders, each with a set of needs, preferences, and constraints, which may be in conflict with those of other stakeholders, thus leading to a negotiations problem.
5. Multiple Objectives: The process COTS selection is carried out with a variety of objectives. These objectives are a result of the need to evaluate products in various non-comparable aspects such as cost, quality, and viability of vendor.

4.5- Determine Project Requirements Elicitation (High level Requirements)

Initial Project's requirements are vital because it defines not only the project product, but also it supports the estimation of a project's budget, resources, and schedule. Resources of initial project requirements are preliminary ideas, documents providing a complete overview of the project components and functionalities are written. Such documentation, also called informal requirements,

consists of natural language documents, possibly garnished with tables, algorithms, finite state machines, and even sometimes implementation considerations [10].

4.6- Project Process Management

Project process management includes methods and tools to undertake the main project processes and activities; initiating, planning, executing, control, and project closure. It interested with assigning the following;

- 1- Identify the project team organization
- 2- Identify project standard and procedures
- 3- Identify project Change management
- 4- Choose the project development methodology

Any defect within the project process management creates a great threat for IS project success [30].

4.7-Develop Statement of Work (SOW)

SOW is a formal document captures and defines the work activities, deliverables, timeline and time estimates for IS project [16]. It is a narrative description of the deliverables or services required to fulfill a contract, this is accompanied by specifies project requirements in a high-level [30], it comprises, Scope of project, Location, Period of Performance, Reporting, Deliverables Schedule, Applicable standards, and Acceptance criteria.

4.8- Project Work Breakdown Structure (WBS)

Work breakdown structure (WBS), is the process of breakdown a project into manageable chunks that can be effectively estimated and supervised in a way that helps organize and define the total work scope of the project [33], WBS consists of three different structures "Project Control Work Breakdown Structure" (PCWBS), "Functional Work Breakdown Structure" (FWBS), and "Relational Work Breakdown Structure" (RWBS) [9].

4.9- Project Cost Estimation

Permanently, cost estimations play a critical role to influence success or failure of the project, Constructive Cost Model (COCOMO) is the famous analytical method to estimate the project cost estimation which uses parameters that are derived from initial projects [42], Estimates of a project's budget, resources, and schedule which represent the outputs need to be updated periodically as more and more information about a project's cost-drivers which represent the inputs [23]. There are four basic steps in project cost estimation; Estimate the size of the development product; which ends up in either Lines of Code (LOC) or Function Points (FP), Estimate the effort in person-months or person-hours, Estimate the schedule in calendar months and Estimate the project cost in dollars (or local currency).

4.10- Project Schedule

Scheduling and sequencing is concerned with the optimal allocation of scarce resources to activities over time, main project scheduling problem involves the scheduling of project activities subject to precedence and/or resource constraints [15], Program evaluation and review technique (PERT) and critical path methods (CPM) techniques are very common and widely adopted management tool being used in project schedule [27], main outlines of the project schedule includes:-

- Identify all the resources that will be necessary to the project and select them in a way that they will be available to perform the projects activities
- Estimate time and resources required for project tasks and activities to complete each task
- Organize tasks concurrently to make optimal use of workforce
- Minimize task dependencies to avoid delays caused by one task waiting for another to complete

4.11-Risk Management

Risk management, is the process of identifying and controlling any activity or event that can negatively influence the IS project success, wherever the risks are all issues that may affect the delivery of benefits to be gained from the IS projects [6]. A Systematic process of risk management is normally divided into three activities Risk identification, Risk analysis, and Risk reduction [25]. In table (3) Han and Huang summarized the main problems with its risk dimensions which cause the project failure when assessing the project Risks management [13].

Table: 3, Main dimensions with their probability Risks of IS projects (Source: Han and Huang, 2007)

Risk Dimension	IS project risk
Users	Resistant to change Conflict between users Users with negative attitudes toward the project Users not committed to the project Lack of cooperation from users
Requirement	Continually changing system requirements System requirements not adequately identified Unclear system requirements Incorrect system requirements
Project complexity	Project involved the use of new technology High level of technical complexity Immature technology Project involves use of technology that has not been used in prior projects
Planning & Control	Lack of an effective project management methodology Project progress not monitored closely enough Inadequate estimation of required resources Poor project planning Project milestones not clearly defined Inexperienced project manager Ineffective communication
Team	Inexperienced team members Inadequately trained development team members Team members lack specialized skills required by the project
Organizational environment	Change in organizational management during the project Corporate politics with negative effect on project Unstable organizational environment Organization undergoing restructuring during the project

4.12-Project Contract

Project contract is an agreement between two or more parties to create mutual business relations or legal obligations. It defines a set of activities to be performed by different parties satisfying a set of terms and conditions [18]. IS Project Contract Model decides the relations among IS project owners,

contractors, managers/consultants and subcontractors, It should be considered that, IS project contract model should be designed to meet the following requirements [41].

1. Requirements regarding the characteristics of IS project and the owner's situation.
2. Fulfill the aim of IS project construction and IS project maintenance that are inseparable.
3. Requirement of IS project risk management.
4. Fulfill the interface management during IS project construction.

4.13- Develop Communication Plan

Efficient and clear communication builds trust between different parties and participants of IS projects that lead to improved collaboration, better team spirit and eventually better results [17]. Communication needs to be written and verbal, and that it needs to be formal and informal, Communication plan takes into consideration all the dimensions of the communication between the project participants. The contents of the plan could include the communication inside and outside the project environment, risks in communication, and how to prepare for them. Also the methods and interval of communication should be included in the plan, e.g. weekly meetings. At the same time the communications plan should include work authorization and deliverable approval methods [34].

4.14- Project Charter

Project charter is a simple, signed agreement between a project team and a representative for those who will use the output (customer). A charter forces the project team and champion to develop a common understanding about the project and creates enough visibility that if a project approach is not sound, the project can be cancelled before it progresses any further [34]. Their format varies from organization to another, but generally it contains a brief scope statement describing the project, how the project supports the strategic goals of the organization, who the project manager is, and finally, if possible, the project's priority within the company. A good project charter should include clear objectives and measurable goals, as well as important assumptions and constraints to guide the project, the lack of a clear project charter increases uncertainty (lack of information), ambiguity (the absence of clarity), and confusion in a project [21].

4.15- Project plan

Project plan is a formal and approved document (based on the project requirements and the established estimates), used to manage and control the execution of the project. It describes the work to be done, the resources required, the methods to be used, the procedures to be followed, the schedules to be met, and the way that the project will be organized. It is a road map to how the project will progress through all project phases and outlined objectives, metrics, and milestones, It includes the follow Specific Practices; [30], [4].

- Document the assumptions, constraints, and alternatives
- Set project scope, cost, and schedule baselines for progress measurement and control
- Provide a tool to communicate with stakeholders
- Establish project milestones and deliverables
- Establish the budget and schedule.
- Identify project risks.
- Plan for data management and project resources.
- Identify functional responsibilities.
- Work Breakdown structure (WBS)
- Plan for needed knowledge and skills.
- Plan stakeholder involvement.

After investigating the critical pre-development activities, it can prove the relation between them and the failure reasons of IS projects. As shown in Table (4) the failure reasons are classified according to each related pre-development activity as evidential correlation. This consequence is based on the statistics in tables (1), (2) and the common failure reasons that listed in the previous researches [13], [14], [32], [1], [5], [18]. This correlation confirms, the most of project failure reasons are attributed and derived from the pre-development activities which deserve our interested. Also, there is a necessity to highlight and focus on the pre-development activities as focal points to decrease the failure probability of IS projects; this can be attained if these activities are given the sufficient time and efforts to be completed in accurate form. Surely, interested in the pre-development activities will contribute to improve the IS projects success.

Table: 4, Correlation between the pre-development activities and the failure reasons of IS projects

Pre-development Activities	Related Failure Reasons
System Service Request (SSR)	Poor project specification Project problem is not clarified
Project Scope Definition.	A3, C10, C14, Scope, objectives unclear Failure to manage changes
Project Feasibility.	A9,C6, Projects are started for the wrong reasons
Custom-building or Commercial Off-The-Shelf (COTS)	Software integration problem Wrong evaluation and selection of COTS
Project Requirements Elicitation	A1, A5, A10, C18.
Project process management	A2, A4, B3,B6,C9,C11,C12, Poor planning processes
Develop a Statement of Work (SOW).	SOW doesn't accordance the extend of the project scope Poor project documentation Insufficient resources.
Work Breakdown Structure (WBS).	A8, C16, Conflict between departments Lack of required team knowledge / skills
Project Cost Estimation	B5,C2, C4, Poor estimation efforts Project under-estimated
Project Schedule	A7,B4,C1,C7, C13,C15, Too long or unrealistic schedules
Project Risk Management.	C3,C8, Risks were not re-assessed, controlled, or managed through the project.
Project Contract	C5, Poor contract management Deadlines and resources not adjusted accordingly
Develop a Communication Plan.	B1, Failure to communicate and act as a team Lack of proper communication
Develop a Project Charter.	Unclear charter for project Failure to adequately identify, project document, and track requirements.
Develop a Project Plan.	A6,B2,C17, Risks not incorporated into the project plan Failure to manage to the project plan Inaccurate project plan.

5- Case study, Virtual Case File Project for the United States Federal Bureau of Investigation (FBI)

The Virtual Case File had been delivered by Science Application International Corporation [8] and it was aimed at facilitating case file management by integrating data from older system, including the Automated Case support system, and eventually replacing them. The FBI had admitted that, the Virtual Case File Technology project had failed to meet the bureau's requirements and that:

- Five years of development and
- \$US 170 million in cost had been lost (National Research Council, 2004)

Case study analysis

At the report of The National Research Council [26] "Review of the FBI's Trilogy Information Technology Modernization Program", The National Research Council saw no evidence that backup and contingency plans had been formalized, the transition plan did not include the availability of the Automated Case Support system after the cut over to the Virtual Case File. Also it found that the requirements for the FBI mission were not included in the Virtual Case File design. Science Application International Corporation said it delivered the first phase of the project ahead of schedule and under budget, but the requirements for the software changed more than one time after the September 11, at the USA. The office of the inspector general found that FBI was still defining the requirements after two years since the start of the project. It also said that the communication with FBI was difficult because of the high turnover of top IT managers.

Through examining the previous project, it can determine the main reasons that cause the project failure and prevent FBI requirements to be met.

- 1- Backup and contingency plans had been formalized
- 2- The transition plan did not include the availability of the Automated Case after cut over
- 3- Requirements for the FBI mission were not completed
- 4- Requirements for the software changed more than one time
- 5- At the first phase of project ahead of schedule and under budget
- 6- FBI was still defining the requirements after two years since the start of the project
- 7- Communication with FBI was difficult because of the high turnover of top IT managers

In the fact, the previous failure reasons are attributed to the following pre-development activities in table (5) this confirms the research objective, the necessity of giving more concern and focus to the pre-development activities which are the main source of projects failure. If these activities are completed accurately, the percentages of IS projects success will be improved.

Table: 5, Failure reasons with the attributed pre-development activity of FBI case study

Failure Reasons	Pre-development Activity
Backup and contingency plans had been formalized	Develop Project Plan Project Risk Management
The transition plan did not include the availability of the Automated Case	Develop Project Plan
Requirements for the FBI mission were not completed	Project Requirements Elicitation
Requirements for the software changed more than one time	Project Requirements Elicitation
At the first phase of project ahead of schedule and under budget	Project Cost Estimation Project schedule
FBI was still defining the requirements after two years	Project Requirements Elicitation
Communication was difficult because of the high turnover of top IT managers	Develop Communication Plan

6- Avoidance Strategy as Solution to Decrease IS Projects Failure

In the context of introducing solutions contribute to decrease the projects failure rates, an avoidance strategy is adopted as a swift solution to avoid the project failure defects through installing a set of corrective guidelines and controls. The purposed strategy is based on many critical success factors that must be considered and embedded in each pre-development activity during the project development execution. Table (6) explains the avoidance strategy to enhance IS project success through eliminating the faults within each pre-development activity and comprises the following guidelines:-

Table: 6, Avoidance strategy' factors to enhance IS project success

Pre-development Activities	Avoidance Strategy Factors
System Service Request (SSR)	- Objectives of the request must be clarified completely - SSR must be described in details - Change in request should be authorized from the head of the department - Problem for the request system must be clarified - Sufficient analysis and study of the request should be performed - Head of the department should signing and issue the SSR
Project Scope Definition	- The project definition must be accurate and completed - Identify clearly the project boundaries and edges - Project goals and objectives must be described completely - Changes in user expectations should be controlled and expected - Scope must be defined completely. - Take care for the project scope creep - Get users involved as early as possible - Found an opportunity to reevaluate or adjust the scope of the project.
Project Feasibility	- Assure that the projects start upon the right reasons and needs - The project costs must accordance the budget. - Estimate the gap between technology and ability - Eliminate poor competencies (and skill shortages) - Avoid the higher cost of capital with existing the ability to provide investment capital - Perfect use of financial resources with proper financial management - The margin between the costs and benefits should be explicitly identified - Calculate the qualitative costs and benefits when evaluate the economic feasibility - Ability to determine the technical and operational needs for the projects - Don't afraid from project cancelation, if it isn't viable
Commercial Off-The-Shelf (COTS)	- Certainty about the requirements of the software under development, about the correctness of the available information, and about the impact of the COTS on the quality of the system. - Different decisions are required to be made at different stages of COTS selection process because COTS selection decisions are multileveled - Integrate COTS software component into software system - Coordinate Multiple Stakeholders, each with a set of needs, preferences, and constraints, which may be in conflict with those of other stakeholders. - Balance Multiple Objectives: The process COTS selection is carried out with a variety of objectives. These objectives are a result of the need to evaluate products in various non-comparable aspects such as cost, quality, and viability of vendor.
Project Requirements Elicitation	- Eliminate Inadequate or vague requirements - Users should be involved with the managing of their expectations. - Don't Change the system to support critical new requirements discovered during final development

	- Assure that no wrong business requirements have been addressed - proper defined of system requirements - Understanding of the main project's requirements completely - Make Project Sponsors write their own requirements
Project Process Management	- It's important to select the appropriate development methodology - Use of mature technology - Abundant of executive support with good project management - Exist the ability to handle the project's complexity - avoid Inadequate or misused methods - setting proper development practices - setting change management - Don't Use a specific development methodology
Develop Statement of Work (SOW)	- Adequate reporting of the project's status - Clear objectives of the project must be stated - All changes in SOW must be performed upon mutually agree - SOW must accordance the extend of the project scope
Project Work Breakdown Structure (WBS)	- Setting a clear WBS Dictionary - Included All tasks and activities in WBS - Flexibility of WBS decomposition and update - WBS must accordance with other structured lists - Eliminate Interrelated conflict between tasks and activities - Consistent between WBS and the project activities
Project Cost Estimation	- Use more than one cost estimation methodology to verify the accurate of estimated cost - The organization has a historical database for organizing and retaining the estimation data - Assure the estimation of supplemental project activities - Allow enough time to do a proper project estimate - Included the cost of project risks into cost estimation - Tasks to be estimated are clearly identified - Estimate the project efforts according to accurate sizing - As much as possible Use at least one software estimation tool - Building cost estimation upon gathering sufficient requirements
Project Schedule	- Fit estimation of time and resources - Don't Using linear approximation when estimating schedule - Estimates for schedule must be correct - Eliminate Unrealistic/Long timescales for the project schedule - Decrease the Pressure of fast delivery to avoid missed deadline and inferior results - Project should have adequate staff to meet the schedule - Optimistic schedules and budgets - Assuming that the time on task equals duration with interruptions
Risk Management	- Project managers must have the requisite experience to be aware of all relevant risk factors - Eliminate conflict between project department and functions - Depend on qualitative and quantitative risk analysis techniques - Accurate risk identification which is the most critical step in risk management. - Effective use of strategies for recruiting and retaining specialized technical personnel for the project staff - Awareness with the system software specifications - Good expectation of risk probability and estimation - Apply suitable risk response strategy - continue to risk evaluation during the project life phases
Project Contract	- Projects should be launched with an agreed contractual completion date - Capability of contractor key person and Effective contract management

	- If Projects take add requirements, deadlines and resources often should be adjusted accordingly.
- Develop Communication Plan	- Eliminate communication breakdowns - Coordinate communication among customers, developers, and users - Adequate communication, including progress tracking and reporting - proper team collaboration & communication - Make delegation and decision making
Project Charter	- Clear charter for project - Sufficient planning - Accessible and relevant documentation for the project - Using simple and flexible documentation - Complete the project documentation - Adequate definitions of roles and responsibilities
Project plan	- Risk should be incorporated into the project plan - Metrics are sufficient to assess the quality of software. - Proper plans and planning processes - Contingency plans for handling problems that arise during the project development - Adequacy of plan and specifications - Strong project plan using documentation support - A project plan should be in sequence form - Complete constructed and just enough time and effort spent on project planning

7. Conclusions and Future Work

An exploration study was conducted for the purpose of surveying the pre-development activities which begin with the System Service Request (SSR) to initiate the projects and end with developing the project plan. Also, investigating the relation between the pre-development activities and the failure reasons of IS projects. Lots of results are concluded in this research (1) there is no IS project suffered from just one failure reason, but all projects had multiple problems and the failure reasons are related to each other, (2) the majority of all project failure reasons are derived from the defects of the pre-development activities (3) Pre-development activities can play an important role to enhance IS project success if these activities executed in proper aspects with sufficient time and efforts to be completed accurately.

Furthermore, this research introduces an avoidance strategy which includes many critical success factors that must be considered for each pre-development activity during the project execution, as a swift solution to decrease the IS projects failure rates. Without doubt, in the case of giving the pre-development activities the reasonable interest, this will contribute to improve the IS projects success.

Governance of the pre-development activities is the future work to enhance the probability of IS projects success and decreases the projects failure rate. It can present many benefits in the area of IS projects, these benefits reflect the necessity of adopting such of this technique as a solution for the failures of IS projects. Governance the pre-development activities shouldn't be restricted in embedded the governance controls only such as (PRINCE2, COBIT, 6-Sigma, and CMMI) but it must be incorporated with installing the mixture of governance mechanisms such as (Arrangements, Structures, Best practices, Metrics, Strategies, Controls, Rules, Relationship, Protocols, Architectures, Guidelines and Standards).

8. References:

1. Al-Ahmad Walid, Al-Fagih Khalid, Khanfar Khalid, Alsamara Khalid, Abuleil Saleem, Abu-Salem Hani, "A Taxonomy of an IT Project Failure: Root Causes" *International Management Review*, Vol. 5, No. 1, pp:93-104, 2009.
2. Bachy G., Hameri A., "What to be implemented at the early stage of a large-scale project" *International Journal of Project Management*, Vol., 15, Issue 4, pp: 211-18, 1997.
3. Barclay Corlane, Osei-Bryson Kweku-Muata "Project performance development framework: An approach for developing performance criteria & measures for information systems (IS) projects" *International Journal of Production Economics* Vol.,124 pp: 272-292, 2010.
4. Calvo-Manzano J. A., Cuevas G., Mejia J., Munoz M. A., San Feliu T., "How is CMMI-DEV applying when using TSPi project planning" *IEEE, Electronics, Robotics and Automotive Mechanics Conference*, Digital Object Identifier:10.1109/CERMA.2009.53, pp:143 – 148, 2009.
5. Cerpa Narciso , Verner M. June " Why Did Your Project Fail?" *Communications of the ACM*, vol., 52, No. 12, pp: 130-134, 2009.
6. Elkington, P., Smallman, C., "Managing project risks: a case study from the utilities sector" *International Journal of Project Management*", Vol., 20, pp: 49-57, 2002.
7. Finnveden Goran, Hauschild Z.Michael, and Ekvall Tomas, "Recent developments in Life Cycle Assessment" *Journal of Environmental Management*, Vol., 91, Issue 1, pp: 1-21, 2009.
8. Friden, Terry, "FBI wasted millions on Virtual Case File" *CNN Washington Bureau*, 4 February, 2005, <http://www.cnn.com/2005/US/02/03/fbi.computers>
9. Golpayegani S. Alireza Hashemi, Emamizadeh Bahram, "Designing Work Breakdown Structures Using Modular Neural Networks" *Decision Support Systems Journal*, Vol., 44, pp: 202-222, 2007.
10. Gorse N., Belanger P., Chureau A., Aboulhamid E.M., Savaria Y., "A high-level requirements engineering methodology for electronic system-level design" *Journal of Computers and Electrical Engineering* Vol.,33 pp: 249-268, 2007.
11. Greer D., Conradi R., "Software project initiation and planning – an empirical study" *Journal of IET Software*, Vol. 3, Iss., 5, pp: 356-368, 2009.
12. Griffith A. F, Gibson G. E., "Alignment During Pre-project Planning" *Journal of Management in Engineering* ,Vol. 17, No. 2, pp: 69-76, 2001.
13. Han Wen-Ming, Huang Sun-Jen, "An empirical analysis of risk components and performance on software projects" *The Journal of Systems and Software*, Vol., 80, pp: 42-50, 2007.
14. Hartman, F., Ashrafi, R.A., "Development of the SMART Project Planning framework", *International Journal of Project Management* Vol., 22, pp:499-510, 2004.
15. Herroelen Willy, "Project Scheduling—Theory and Practice" *Production and Operations Management Journal*, Vol., 14, No., 4, pp: 413-432, 2005.
16. Jeffrey A. Hoffer, Joey F. George, Joseph S. Valacich, "Modern Systems Analysis and Design", Prentice Hall, ISBN-10: 0132240769, pp: 45-125, 2007.
17. Katainen Tommi, Nahar Nazmun, "Using Methods and IT Tools Innovatively for the Management of International IS Development Projects" *IEEE, Portland International Conference Management of Engineering & Technology*, pp:1851-1863, 2008.
18. Krishna P. Radha, Karlapalem K., Chiu D.K.W., "An ER framework for e-contract modeling, enactment and monitoring" *journal of Data & Knowledge Engineering* vol., 51, pp: 31-58, 2004.

19. Lepage Darren "Driving Success in Substation Control and Information System Projects" Utility Automation and Engineering journal, pp:34-42, February 2009, www.utility-automation.com.
20. Mahaney C. Robert, Lederer L. Albert, "The role of monitoring and shirking in information systems project management" International Journal of Project Management, Vol., 28 pp: 14-25, 2010.
21. Mahring Magnus, Keil Mark, "Information Technology Project Escalation: A Process Model" Decision Sciences journal, Vol., 39, No., 2, pp:239-272, 2008.
22. Mahring Magnus, Mathiassen Lars, Keil Mark, Pries-Heje Jan "Making IT Project De-Escalation Happen: An Exploration into Key Roles" Journal of the Association for Information Systems, Vol. 9 Issue 8, pp:463 462-496, 2008.
23. Malik A. Afzal, Boehm Barry "Quantifying requirements elaboration to improve early software cost estimation" Information Sciences Journal, <http://dx.doi.org/10.1016/j.ins.2009.12.002>
24. Malkinson J. Terrance "A "toolkit" for information technology proposal and project documentation" IEEE, Professional Communication Conference, IPCC 2001, pp: 205-220, 2001.
25. Mojtahedi S. Mohammad, Mousavi S. Meysam, Makui Ahmad "Project Risk Identification and Assessment Simultaneously Using Multi-Attribute Group Decision Making Technique" Safety Science Journal, Vol., 48, pp: 499-507, 2010.
26. National Research Council, "A Review of the FBI's Trilogy Information Technology Modernization Program", Computer Science and Telecommunication Board, National Academies Press, Washington D.C., 2004, <http://www.nationalacademies.org/nrc/>.
27. Omar Anwar, "Uncertainty in Project Scheduling, Its Use in PERT/CPM Conventional Techniques" Cost Engineering Journal, Vol. 51, No. 7, pp: 30-34, 2009.
28. Pan Gary, Hackney Ray, Pan Shan "Information Systems implementation failure: Insights from prism" International Journal of Information Management, Vol., 28, pp: 259-269, 2008.
29. PM Hut (Project Management Hut), "Project Failure Statistics and Facts", <http://www.pmihut.com/project-failure-statistics-and-facts>, <http://www.standishgroup.com/>, September, 2008,
30. Project Management Institute "A guide to the Project Management Body of Knowledge PMBOK Guide", An American National Standard, Fourth Edition, PA 19073-3299, 2008.
31. Ram L. Kumar "Managing risks in IT projects: an options perspective", Information & Management Journal, Vol., 40 pp: 63-74, 2002.
32. Rodney A. Stewart, "A framework for the life cycle management of information technology projects: Project IT", International Journal of Project Management, Vol., 26, Issue 2, pp: 203-212, 2008.
33. Scasso R. Heredia, Larenas G. Santana "Project-breakdown structure: the tool for representing the project system in project management", International Journal of Project Management, Vol., 9, Issue 3, pp:157-161, 1991.
34. Tesch Debbie, Kloppenborg Timothy, Frolick Mark "IT Project Risk Factors: The Project Management Professionals Perspective" Journal of Computer Information Systems, pp: 61-69, Summer 2007.

35. Uher E.Thomas, Toakley A. Ray, "Risk management in the conceptual phase of a project" International Journal of Project Management, Vol., 17, Issue 3, pp: 161-169, 1999.
36. Uppal Kul B., "Project Management, Cost Engineering, Project Definition, Action Plans or What?" AACE International Transactions, PM.01.1, 2008.
37. Wanyama Tom, Far B. Homayoun, "Towards Providing Decision Support for COTS Selection" IEEE, Canadian Conference on Electrical and Computer Engineering, pp: 908-911, 2005.
38. Whitten, Jeffrey Bentley L., Lonnie D., "Systems Analysis and Design for the Global Enterprise", MCGRAW HILL, ISBN 9780071107662, 2007.
39. Yan B., Benslimane Y., Yang Z., "IS Development Activities and Methodologies in Practice: A Survey of the IT Sector in China" IEEE, Issue Date: 8-11 Dec., 2009, Print ISBN: 978-1-4244-4869-2, pp: 598 – 602, 2009.
40. Zhang G. Peter, Keil Mark , Rai Arun, Mann Joan, "Predicting information technology project escalation: A neural network approach" European Journal of Operational Research Vol.,146 pp: 115–129, 2003.
41. Zheng Chan, Feng Jingchun, Li Ming, "Research on IT Project Contract Model" IEEE, Second International Conference on Intelligent Computation Technology and Automation, ICICTA, Vol., 2, pp: 172 – 175, 2009.
42. Zheng-Wei Huang, "Cost Estimation of Software Project Development by Using Case-Based Reasoning Technology with Clustering Index Mechanism" IEEE, Fourth International Conference on Innovative Computing Information and Control (ICICIC), pp: 1049 – 1052, 2009.

Arabic Spatial Query Expansion Based on Egypt Geo-Ontology

Alaa Hedib¹

Akram Salah²

Abstract

Until recently most web search engines have treated geographical terminology within user queries in the same way as other terminology, a typical search engine only return web pages that include *placenames* involved, but usually doesn't retrieve other web documents that are associated to that place, although it may be user interest. This paper aim to introduce architecture for using geographical Ontology built for the Egypt geographical features to expand the Arabic users' queries which contain an Egyptian place names.

Keywords: Geographical information retrieval, Ontology, spatial query, queries extraction.

1. Introduction

A large amount of geographical information is currently being stored and delivered over the internet. Geographical Information Retrieval (GIR) aims to develop information retrieval (IR) systems with spatial awareness and using geo-spatial information to improve retrieval effectiveness of finding information resources that relate to particular geographical locations. For example if a user seeks to get information proximity related to somewhere, e.g., "killing crimes near of DEWIK region", such spatial query need a search system aware of spatial relations and capable of utilizing it in search for documents.

¹ Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University.

² Department of Computer Sciences, Faculty of Computer Information, Cairo University.

The relationships reflect the geographical distances may include expressions such as, in, part of, north, south, east, west, and near-of. An intelligent information retrieval system should be able to recognize such expressions. It also should be able to interpret such expressions semantically utilizing a body of knowledge that represent the geographical relationships.

In this paper we introduce a system that is capable of interpreting Arabic queries requesting information about Egyptian places. In this system we are able to interpret queries involving geographical expressions. We utilize an ontology representing Egypt's geographical features.

The paper proceeds as follows. In the next section we start with background for the main concepts related to the spatial query and query expansions. Section 3 overviews some related works. Then section 4 aims to explain a proposed geographical information system structure using the Egypt geo-ontology for expand the spatial Arabic queries. Finally we have some concluding remarks.

2. Background

This section will provide the main concepts which related to the spatial query and the query expansion methods and the uses of the Ontology in the query expansion process.

2.1. Spatial Query vs. Generic Query

A spatial query is different from a generic one in that it usually includes one or more spatial terms, spatial term here refer to a geographical feature . It is often used by a user when he/she wishes to find Web resources that are related to a specific place. If we check a generic syntax of the spatial query we will find it contain these mean three parts <info user seek to know > <relation > <location >.

2.2. Spatial Query Challenges

Spatial query must be handling by custom method, the handling way in conventional search engines only return web pages that include the *placename* involved, but usually do not retrieve the other web documents that are geographically associated to that place, typically that refers to places in the proximity of the subject place or is contained within it, although it may be of

interest. There must be a representation for various places with their interrelationships. The relationships reflect the geographical distance such as north, south, east, west, and near-of. The format must be represented in machine readable format to allow programs to deal with it automatically.

2.3. Query expansion (QE) (query augmentation).

Query expansion, sometimes called query augmentation, is “the process of reformulating a seed query to improve retrieval performance in information retrieval operations.” [Vectomova, 2006]. Another definition of the query expansion is “Is to add new meaningful terms to the initial query.” [J. Bhogal, 2006].

Parsing the above two definitions the query expansion process contains two operations:

- Reformulate some of its terms (correct miss spelling...). and
- Add new terms to the initial query terms. This process of adding terms, take a place as one of the important research point in this field.

2.4. Manual vs. Automatic Query Expansion.

Query expansion can be performed manually, automatically or with user-assisted. Manual query expansion relies on user expertise to make decisions on which terms to include in the new query. Automatic query expansion calculates weightings of all terms. The terms which have the highest weighting are added to the initial query. Retrieval performance depends on how the weightings have been calculated. However, there are some user-assisted query expansion systems which generates possible query expansion terms and the user selects which of these to include.

2.5. Query expansion methods.

Query expansion methods can be either depend on search result or knowledge structure forms. Query expansion based on the search results used relevance feedback and probabilistic. Relevance feedback is a fairly established technique for modification of the initial query using words from top-ranked or identified relevant documents.

Using relevance feedback the query can expand with the new terms by

- *User-assisted:* user selects the more relevant terms from the retrieved document by the initial query. This called *Traditional relevance feedback*.

- *OR Automatic*: automatically select n document from the retrieved documents and use their terms to add to the initial query. This called pseudo relevance feedback, [J. Bhogal ,2006].

The statistical-based query reformulation strategies not efficient with all query types as it work at term level, not entity level, they don't use the semantic information given by such entities which is considerably relevant for some types of queries e.g. (spatial query).

Query expansion based on some forms of knowledge structure. Knowledge base systems used to adapt and represent the information concepts and its relationships. As,

- Using the WordNet :

WordNet, is a lexical database for the English language words, including the synonyms words sets and the semantic relations between these synonyms. So it can used to find the words that are lexically related to the original query words, and use it as expanded terms.

- Using semantic network:

Additional query terms are derived by traversing a semantic network built up according to a knowledge structure. Knowledge structures used by this group of techniques can either be a general-purpose ontology (or thesaurus), or an ontology built for a specific domain. . [Gaihua ,2005].

2.6.Ontology-based Query Expansion:

Ontologies provide a structured way of describing knowledge. According to [Gruber, 1993] Ontology is a “shared specification of a conceptualization”. Ontologies used in Semantic Web to enrich the current Web resources with some well-defined meaning to enhance the search capabilities of existing web searching systems. [Roberto, 2003]

The simply method for using Ontology to expand the query:

- Query terms must be disambiguated so that it map to a unique ontology concept.
- Then terms related in the ontology to the disambiguated concepts are returned and added to the original query.

The method used in this paper for the proposed system, is using the ontology-based query expansion to expand the Arabic spatial query terms that maps to concepts in the Egypt geographical ontology.

3. Related Work

The following projects indicate to the activity that do in the field of working with the geographical information retrieval, however in the place name extraction from the document or expansion the query which has geographical terms, or build geographical gazetteers.

3.1. The European (UK ,France, Germany, Swaziland) Research Project, SPIRIT (SPatially-aware Information Retrieval on the InTernet):

It is a research project that has been engaged in the design and implementation of a search engine to find documents and datasets on the web related to places or regions referred by a specific query. The main components are the user interface, core search engine, the geographical ontology, web document collection, textual and spatial indices of document collection, relevance ranking and metadata extraction. [Abdelmoty , 2004]. Using the user interface user specify the subject of interest, place name and spatial relationship to the place name. The subject term or terms may be non-spatial, as in “database conferences in USA”, or spatial, e.g. “hotels”, or “cities in the UK”. Different types of spatial relationships need to be supported including, topological, e.g. inside, outside, directional, e.g. north-of and south-of and proximal, e.g. near to or within a specified distance of. Then the system takes this query expression and starts a process of query disambiguation and expansion, the place ontology play core roles as it used here for offered the alternative place names and take the combination of the place name and spatial relationship to extent them to the locations of interest. Now these system’s interpretation of the query, it is submitted to the core search engine and a relevance-ranked list of documents is returned to the user. Web documents are structured to facilitate indexing. SPIRIT supports both pure text indexing and spatio-textual indexing. Here we will find another role for the place ontology, In order to support spatial indexing of the document collection, each referenced document that contains place names (which identified by the ontology) is associated with one or more “document footprints” that are derived from the geographical ontology entries for the respective names. The search engine uses the web document collection and its text and spatio-textual indexes to process the query by comparing the “query footprint” and the “documents’ footprints”. The relevance ranking component takes results retrieved from the search engine and relevance ranks

them with respect to the non-spatial and spatial elements of the query. SPIRIT Results listed as URLs Plus symbols on map.

Ontology is the primary component of the SPIRIT which provides a model of the terminology and structure of the geographic space, The geo-ontology plays a key role in the interpretation of user queries it used to recognize the presence of place names in a query as it stores knowledge of instances of place names with alternative names, place types, qualitative spatial relationships to other places and then to perform disambiguation; the formulation of system queries, generation of spatial indexes, relevance ranking and metadata extraction. This geographic ontology is represented in the DAML+OIL language. [ROSS S, 2006].

3.2. GREASE project (Geographic REASONing for Search Engines):

Under this project XLDB Team of the University of Lisbon, Faculty of Sciences develops Geo-Net-PT 02 which is a public Geographic Ontology of Portugal. Geo-Net-PT contains all the geographic administrative data of Portugal, it is encodes the geographic knowledge and provides it to GIR components, this knowledge are: 1) features (an unambiguous location. It can be described by one or more place names.), 2) Feature types (classes of features. For example, island, mountain, city), a feature has only one feature type, 3) and relationships among them. Links joining features OR feature types: part of, adjacent, capital of. [Francisco J, 2009].

Geo-Net-PT02 is published in the Web Ontology Language (OWL).

3.3. Geo-referenced Information Proccession System GIPSY:

It is one of the first geo-referenced information processing systems that utilized a gazetteer for matching spatial terms, GIPSY basic strategy is getting the area which the document refer to , by extraction the place names from the document its algorithm based on a gazetteer or thesaurus containing place names and the names of other geographically significant object. [Woodruff, 1994].

4. An Egyptian Arabic Geographical Information Retrieval system.

Using Ontology based query expansion method in the geographical domain, provides the needed information related to the geo terms expansion. For example it provides the placename synonyms and the value of the spatial relationships such as (in, part-of, east, west, north, south, near-of). Using such

semantically representation for the geographical information with some preprocessing for the query terms can produce an efficient geographical information retrieval system.

In the below graph we try to provide an architecture to handle the Arabic spatial query expansion using the Egyptian geographical Ontology.

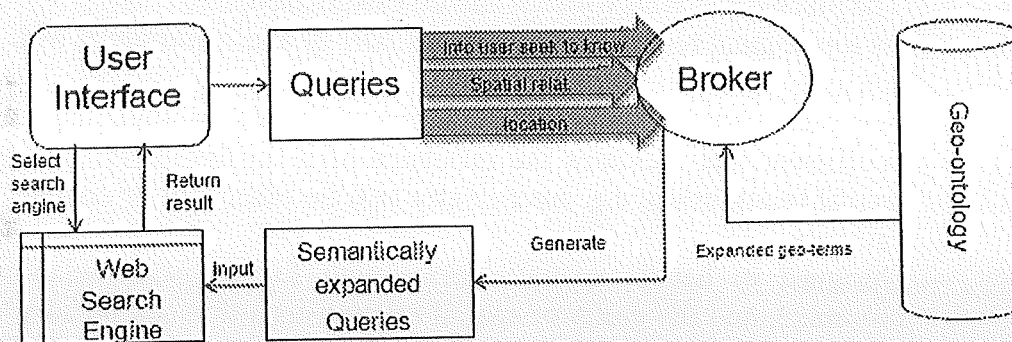


Figure (1): the expansion architecture life cycle.

The Geographical Ontology plays a key role as it provide the needed answers for such type of questions which related to the Egypt locations. For example:

There is the *near-of* spatial relationship used with placename.
There is the *west* spatial relationship used with placename synonym.
There is the *in* spatial relationship used with area.

الاماكن القريبه من الدويقة
الاماكن التي تقع غرب العاصمة
الاماكن الموجوده في صعيد مصر

4.1. The Egypt Geo-Ontology:

In this research we have built Egypt Geo-ontology based on the standard publication by the Egypt State Information Service (SIS)³, which provides the

³ <http://www.sis.gov.eg/En/Story.aspx?sid=6>

Egypt's Administrative Division⁴, we uses the release of 2006. And adding some new higher levels of geographical taxonomy as well as adding relationships among the places.

Several types of information are encoded in the geo-ontology including:

- The various names that a place is known by (synonyms),
- The place types with which it can be categorized,
- Its topological relationships (such as *part-of* and *containing*) with other places,
- For each category of place, the geographical ontology maintains a coefficient that indicates the influence of it on the interpretation of a spatial relationship, e.g. *west-of city* will be other city or another place category.
- Designed to handle spatial queries with fuzzy spatial relationships presented, and the group of spatial relationships that can be handled by the geo-ontology e.g. *in, near, outside, north-of, south-of, east-of, west-of*.
- The availability of the alternative interpretation of a spatial relationship *rel*. For example, for *near*, three options may be available for its interpretation: an area covers only *where*, an area covers both *where* and its surrounding regions and an area covering neighboring regions of *where*. We assume that, the *near rel* will implement as equivalence for the properties (*north-of, south-of, east-of, west-of*).

⁴ <http://www.eip.gov.eg/DataDirectories/AdministrativeDivision.aspx>

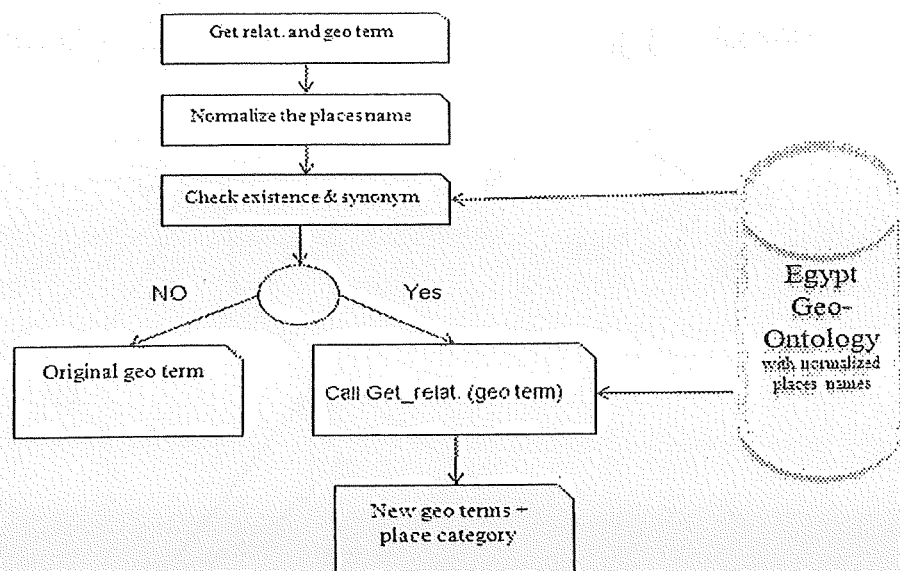


Figure (3): broker roles

4.3. User Interface:

User interface built to allow users to enter their geo-related queries, display the new semantically expanded queries, and allow for user to select one of the most popular the search engines to handle these queries.

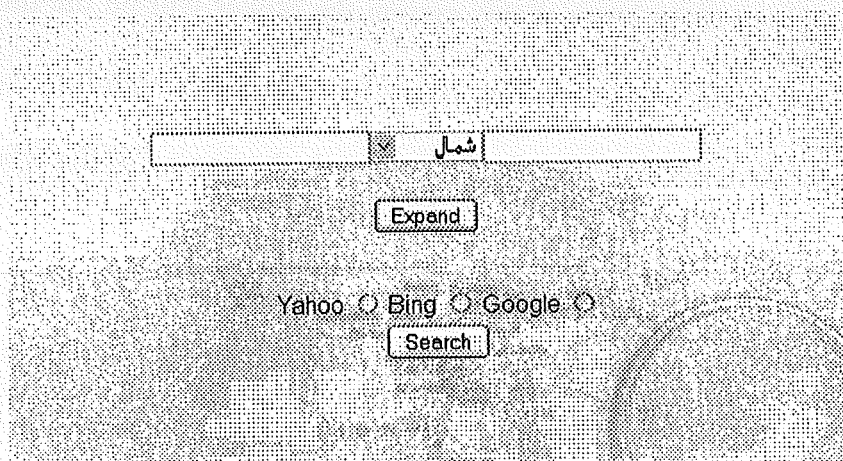


Figure (4): system GUI

4.4. Results:

Table (1): sample results

System semantically expanded query	User query
اهم الاحداث التاريخيه فى (محافظة اسيوط OR محافظة سوهاج OR محافظة اسوان OR محافظة منيا OR محافظة وادى جديد OR محافظة اقصر OR محافظة بنى سويف OR صعيد)	اهم الاحداث التاريخيه فى الصعيد
اسباب الانهيارات الارضية بالقرب من الدويقة OR منشاه ناصر OR منشيه ناصر OR مساكن سوزان مبارك)	اسباب الانهيارات الارضية بالقرب من الدويقة
المشروعات التنمويه غرب العاصمة OR القاهره)	المشروعات التنمويه غرب العاصمة

Table 1, shows sample of Arabic spatial queries contains Egyptian placename or its synonyms with geographical relationships, and the result of new semantically expanded queries. We note that the system interpret the geographical expressions and generate the new terms that can improve the information retrieval process.

5. Experiments and IR performance Comparison.

The TREC-2001 Arabic corpus, also called the AFP_ARB corpus, consists of 383,872 newspaper articles in Arabic from Agence France Presse. This fills up almost a gigabyte in UTF-8 encoding as distributed by the Linguistic Data Consortium. There were 25 and 50 topics used in 2001 and 2002 respectively with relevance judgments, available in Arabic, French, and English, with Title, Description, and Narrative fields. We used the Arabic titles and descriptions as queries of the 8 topics in the experiments, these topics contain Egyptian places names, which we can consider it as spatial queries.

Through the experiment query words normalized, then expanded spatially.

Table (2): experiment results

Q no	Num_ret	Num_rel	Num_rel_ret	Num_rel_ret_spatial
12	1000	24	22	22
13	1000	174	122	120
16	1000	57	5	15
18	1000	321	214	223
19	1000	6	0	0
20	1000	242	4	5
33	1000	3	3	3
34	1000	13	3	4

Table 2 has record for each spatial query, contain, query number inside the corpus topics (Q no), total number of retrieved documents (Num_ret), number of judged relevant documents (num_rel). In the experiment using the query words, (num_rel_ret) represent the numbers of relevant retrieved document, (num_rel_ret_spatial) represent the numbers of relevant retrieved document when apply the spatial expansion.

The results show that, most queries which spatially expanded retrieve more relevant documents.

6. Conclusion & Future work.

Using the Egypt geographical ontology to semantically expand the spatial Arabic query by handling the place name synonyms and the spatial relationships help to improve the information retrieval operation by increasing the *recall* ratio.

In geographic space, "everything is related to everything else", so we can work with the document indexing to get more improvement in the information retrieval operation by increasing the *precision* ratio. There are different multi-dimensional indexes have been proposed for managing spatial data, we will study it and try to find and improve one of the spatial indexing methods to be suitable with our requirement .

Reference

- (Abdelmoty, 2004) Abdelmoty A., Philip D., Christopher B., Gaihua F. and David F., "A critical evaluation of ontology languages for geographic information retrieval on the Internet," School of Computer Science, Cardiff University, Wales, UK, 2004.
- (Christopher, 2005) Christopher B. and Ross P., "GIR'05, 2005 ACM Workshop on Geographical Information Retrieval, Bremen, Germany," School of Computer Science Cardiff University and Department of Geography, University of Zurich, 2005.
- (Gaihua, 2005) Gaihua F., Christopher B. and Abdelmoty A., "Ontology-based Spatial Query Expansion in Information Retrieval," School of Computer Science, Cardiff University, Cardiff, UK, 2005.
- (Gruber, 1993) Gruber T., "A translation approach to portable ontology specifications," Knowledge Systems Laboratory, Stanford University, 1993.
- (Francisco, 2009) Francisco J., Marcirio C., Catarina R. and Mário J., "Geographic Ontologies Production in GREASE-II," Department of Informatics Faculty of Science University of Lisbon, 2009.
- (J. Bhogal, 2007) J. Bhogal b, A. Macfarlane a, P. Smith "A review of ontology based query expansion", School of Informatics, City University, Northampton Square, London EC1V 0HB, United Kingdom b Department of Computing, University of Central England, Feeney Building, Perry Barr, Birmingham, 2007.
- (Eldesouki, 2009) Eldesouki M, Arafa W, Darwish K, "Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective", Institute of Statistical Studies and Research, Cairo University, 2009.
- (Reid J, 2003) Reid J., "geoXwalk – A Gazetteer Server and Service for UK Academia" Koch/Sølvberg (Eds) Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003.

(Roberto, 2003) Roberto N. and Paola V., “An Analysis of Ontology-based Query Expansion Strategies”, Dipartimento di Informatica Università di Roma “La Sapienza”, 2003.

(Ross, 2006) ROSS S. and PAUL C. *The Design and Implementation of SPIRIT: a Spatially-Aware Search Engine for Information Retrieval on the Internet*. Department of Geography, University of Zurich, Switzerland and Department of Information Studies, University of Sheffield, UK and others, 2006.

(Vechtomova, 2006) Vechtomova O. and Wang Y., "A study of the effect of term proximity on query expansion", *Journal of Information Science* 32 (4): 324–333, 2006.
<http://jis.sagepub.com/cgi/content/abstract/32/4/324>. Retrieved 2006-12-09

(Woodruff, 1994) Woodruff A. and Plaunt C., “GIPSY: Automated geographic indexing of text documents”. *Journal of the American Society for Information Science*, 45(9): 645–655, 1994.

USING MODULARITY WITH ROUGH DECISION MODELS

Ahmed T. Shawky, Hesham A. Hefny, and Ashraf H. Abd-Elwahab

ABSTRACT

Decision models which adopt rough set theory have been used effectively in many real world applications. However, rough decision models suffer the high computational complexity when dealing with datasets of huge size. In this research we propose a new rough decision model that allows making decisions based on modularity mechanism. According to the proposed approach, large-size datasets can be divided into arbitrary moderate-size datasets, then a group of rough decision models can be built as separate decision modules. The overall model decision is computed as the consensus decision of all decision modules through some aggregation technique. This approach provides a flexible and a quick way for extracting decision rules of large size information tables using rough decision models.

KEYWORDS

Rough sets, Fuzzy sets, modularity, Data mining.

1. INTRODUCTION

During the last two decades, rough set theory has received much attention as a promising technique for data mining. Its ability to deal with imperfect data analysis makes it quite efficient for drawing useful decisions from datasets of various real world applications. However, rough decision models suffer the problem of high computational complexity of extracting decision rules when dealing with large-size datasets. Several researchers have proposed various approaches to overcome this problem. One of the recent attractive approaches, has been suggested by C. Degang and Z. Suyun [1], to improve the performance of rough decision making process, by integration of fuzzy set and rough set theories. This approach provides two benefits; the first is turning the continuous-valued conditional attributes into nominal or ordinal values which greatly simplifies the computations of finding reducts. The second benefit is the ability to deal with attributes with uncertain values which can be handled with fuzzy linguistic values that differ from decision maker to other. In this paper we make a further step for decreasing the computational cost of finding rough decision rules by introducing the approach of modularity. Modular approach to decision making uses the central idea of task-decomposition to reduce the computational cost of drawing decisions over large datasets. Modular neural networks are successful applications of such an approach [2, 3]. In this paper, we show how this approach can be adopted to reduce the computational cost of large rough decision models. This paper is organized as follows: section 2 presents the approach of modularity, section 3 discusses rough sets in details, the proposed modular rough decision model (MRDM) is presented in section 4, and section 5 concludes the paper.

2. THE APPROACH OF MODULARITY

Modular approach for decision making has been firstly proposed in literature in the works of R. A. Jordan et al (1991, 1994), see [4, 5]. The basic idea behind modular approach is task-decomposition, where large tasks can be divided into relatively small ones to be handled easily. Similarly, Large decision making tasks can be decomposed into small ones among group of local experts to reduce the cost of the overall decision. Modular design has been successfully used in various areas, e.g., robotics and neural networks as presented by Tseng, and Almogahed [2].

Procedure in modular design starts with decomposing the given system (task) into subsystems, i.e. modules, for simpler design, followed by aggregating the modular designs. The idea is to ignore interconnection among these subsystems in the design stage. Since the subsystems are smaller than the original system, the design effort and computation needed in each subsystem design are typically lower. The system will also be easier to debug and maintain due to a smaller system size. In many cases, appropriate decomposition of modules is a designer's issue. Figure (1) represents modular expert with k local experts with outputs $O_1, O_2, \dots, O_k$, while $g_1, g_2, \dots, g_k$ are the integrating weights, and Σ denotes to gating process.

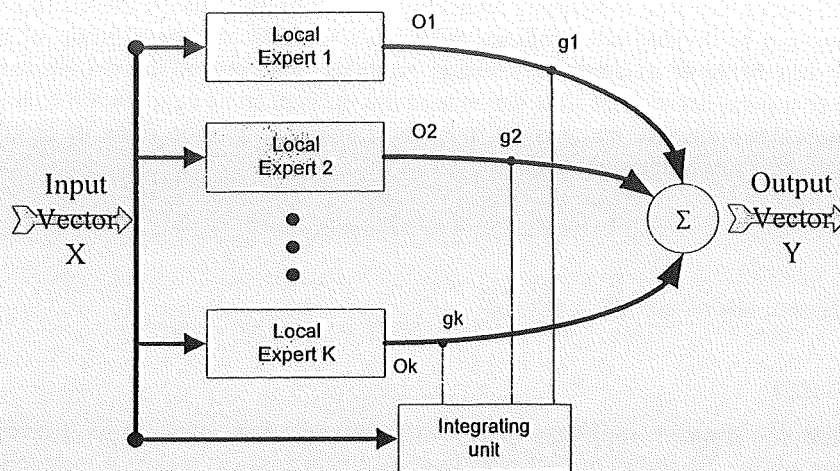


Figure 1. Modular expert architecture

2.1 ELEMENTS OF MODULARITY

When considering modular structure to solve a problem, one has to take into account the following points by P. Melin et al [3]:

- Decompose the main problem into subtasks.
- Organizing the modular architecture, taking into account the nature of each subtask.
- Communication between modules is important, not only in the input of the system but also in the response integration.

2.2 RESPONSE INTEGRATION

Integrating the response outputs of decision modules can be one of the following choices [3]:

- The method of "Winner Takes All", for problems that require similar tasks.
- Models in series, the output of one module is the input to the following one.
- Voting methods, for example the use of the "Softmax" function.
- Linear combination of output results.
- Using discrete logic.
- Using statistical methods.
- Using fuzzy logic.

3. ROUGH SETS

Rough set theory, proposed by Z. Pawlak (1982), can serve as a mathematical tool for dealing with classification of datasets based on uncertainty concepts [6, 7]. Various successful rough decision models have been developed in several fields [8]. The following subsections give a more detail discussion about the main characteristics of rough set theory.

3.1. Information Systems

According to Pawlak (1982), Data are usually presented in the form of decision tables, also called information systems, which consist of rows and columns. Rows of the decision table represent cases, while columns represent variables. Such variables are considered as properties of each case. The set of independent variables are called conditional attributes and a dependent variable is called a decision attribute. Table 1 shows an example of a decision table.

Table 1. Example of Decision table.

Object	Conditional Attributes			Decision Attribute
	Age	Height	Gender	Accepted
X1	Young	Tall	Male	Yes
X2	Baby	Tall	Female	Yes
X3	Young	Tall	Female	Yes
X4	Old	Medium	Female	No
X5	Baby	Short	Male	Yes
X6	Old	Medium	Male	No

An information system S , as presented by Y. Qian et al [9], is a pair (U, A) , where U is a non-empty, finite set of objects and is called the universe and A is a non-empty, finite set of attributes. V is the set of all attribute values, such as $V_a: U \times A \rightarrow V$ for each $x \in U$. According to the example shown in Table 1, $U = \{X1, X2, X3, X4, X5, X6\}$, $A = \{\text{Age, Height, Gender, Accepted}\}$, and $V(X1, \text{Age}) = \text{Young}$.

3.2. Indiscernibility Relation

One of the fundamental ideas of rough set theory is the Indiscernibility relation. For $B \subseteq A$ and $x, y \in U$, the Indiscernibility relation $\text{IND}(B)$ is a relation on U defined as follows [10]:

$(x, y) \in \text{IND}(B)$ if and only if $V(x, a) = V(y, a)$ for all $a \in B$. The Indiscernibility relation $\text{IND}(B)$ is an equivalence relation. Equivalence classes of $\text{IND}(B)$ are called elementary sets and are denoted by $[x]_B$. Elementary sets may be computed by using attribute-value pair blocks. Let $a \in A$ and let v be a value of a , for some case. For complete decision tables if $t = (a, v)$ is an attribute value pair, then a block of t , denoted $[t]$, is a set of all cases from U that for attribute a has value v . For example if $t = (\text{Gender, Male})$ then $[t] = \{X1, X5, X6\}$.

3.3. Reducts

For $B \subseteq A$, the corresponding partition on U will be denoted by B^* . B is called reduct if and only if [3]: $B^* = A^*$ and, B is minimal if: $(B - \{a\})^* \neq A^*$ for all $a \in B$

For example, from Table 1 we have: $A^* = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$

Let $B = \{\text{Age, Height, Gender}\}$, and $C = \{\text{Age, Gender}\}$

Then, it is clear that: $B^* = A^*$, and $C^* = A^*$

Also $(B - \{\text{Height}\})^* = A^*$ this means that B is not minimal

On the other hand, C is reduct of A since

$(C - \{a\})^* \neq A^*$ for all $a \in C$.

3.4. Approximation space

In a completely specified decision table, any finite union of elementary sets, associated with B, will be called B-definable set $[x]_B$. The concept of "approximate space" has arisen due to the fact that not any subset of X is B-definable [10, 11]. In this case, the concepts of lower and upper approximations are defined on the basis of the Indiscernibility relation. Let X be any subset of the set U of all cases. The set X is called a "concept" and is usually defined as the set of all cases defined by a specific value of the decision attribute. In general, X is not a B-definable set. However, set X may be approximated by two B-definable sets; the B-lower approximation of X, denoted by $\underline{B}X$ and the B-upper approximation of X, denoted by $\overline{B}X$ which are defined as follows:

$\underline{B}X = \{x \in U \mid [x]_B \subseteq X\}$, and

$\overline{B}X = \{x \in U \mid [x]_B \cap X \neq \emptyset\}$

The above shown way of computing lower and upper approximations, by constructing these approximations from singletons x, will be called the first method. The B-lower approximation of X is the greatest B-definable set, contained in X. The B-upper approximation of X is the smallest B-definable set containing X. This concept can be illustrated from Table 1 as follows:

let $U = \{X1, X2, X3, X4, X5, X6\}$ and $A = \{\text{Age, Height, Gender, Accepted}\}$

If $B \subseteq A$ and $B = \{\text{Height}\}$, then $B^* = \{\{X1, X2, X3\}, \{X5\}, \{X4, X6\}\}$

Suppose we have a new concept $X = \{X2, X3, X5\}$, in this case, we have

Lower approximation $\underline{B}X = \{X5\}$

Upper approximation $\overline{B}X = \{X1, X2, X3, X5\}$

Once the concepts of lower and upper approximations are adopted, then we can distinguish three regions in approximation space as follows:

- The positive region $\text{POS}(BX) = \underline{B}X$
- The boundary region $\text{BND}(BX) = \overline{B}X - \underline{B}X$
- The negative region $\text{NEG}(BX) = U - \overline{B}X$

3.5. Rule Induction

For the inconsistent input data, Y. Qian et al [11] and J. W. Grzymala-Busse and S. Siddhaye [8] explained that the rules induced from the lower approximation of the concept certainly describe the concept, so they are called certain. On the other hand, rules induced from the upper approximation of the concept describe the concept only possibly (or plausibly), so they are called possible.

For example: As a certain we can say:

If (Age, old) and (Height, medium) then (accepted, no)

As a possible we can say:

If (Gender, Female) then (accepted, yes) with $\alpha = 0.67$

α is called a confidence factor and can be defined as the percentage of the number of elements that are in the elementary set and satisfy the concept for the rule from the total number of elements in the elementary set (upper approximation) in this example

$B = \{\text{Gender}\}$ then $B^* = \{\{X1, X5, X6\}, \{X2, X3, X4\}\}$

$X = \{X2, X3\}$ $P|X| = 2$

$\bar{B}X = \{X2, X3, X4\}$ $P|\bar{B}X| = 3$

$$\alpha = \frac{P|X|}{P|\bar{B}X|} \quad \alpha = \frac{2}{3} = 0.67$$

4. MODULAR ROUGH DECISION MODEL (MRDM)

The proposed MRDM aims to simplify the process of drawing decisions using rough decision model over a large information system. This is can be performed through what we call "Grid Modular" approach, through which the given large information system is split into a group of moderate size sub-information systems. Then, each sub-information system is treated as a separate rough decision module. Through the MRDM proposed model, we can control the number of modules created from the main information system. The final decision of the overall decision model is computed through a gating technique of the output decisions of all decision modules. In this paper, we choose voting as the gating technique of the proposed MRDM. Figure 2 represents the structure of proposed model.

The main adopted algorithms within through the proposed MRDM are as follows:

- Algorithms 1: Which returns the confidence degree of the considered association rule based on the information table as a whole.
- Algorithm 2: Which splits the original information system into n-sequential module.
- Algorithm 3: Which splits the original information system into n-random module.
- Algorithm 4: which performs the gating of the output decisions of n-decision modules.

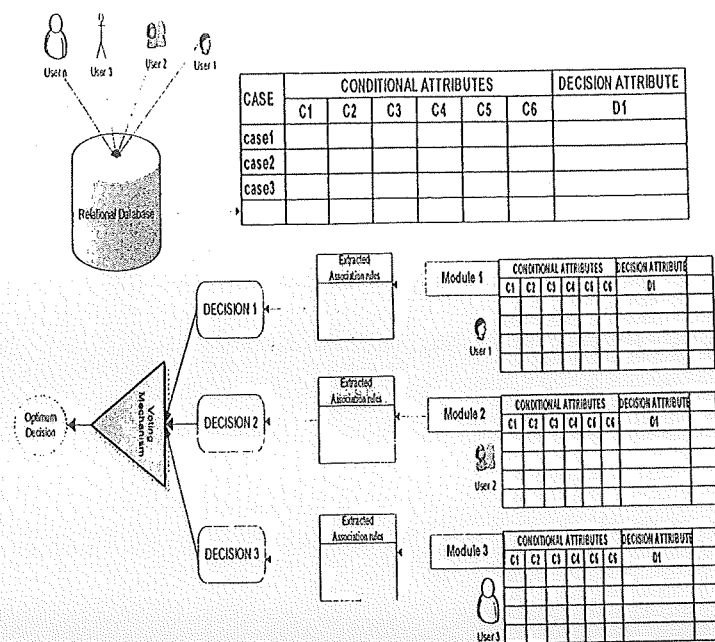


Figure 2. Structure of the Proposed MRDM

Algorithm (1) Returning a decision from the Information Table as a whole

// Input: rough.table(U,A) is an information table
// Output: chosen decision, greatest α degree

```

1  Read Rule, nofdecisions
2  Define BX, TX, k, l, X[], alfa[], Descision[]
3  BX = |cases check with Rule|
4  For i = 0 to nofdecisions
5      Input Descision[i]
6      X[i] = |cases check with Rule ^ cases which satisfy Descision[i]|
7      alfa[i] = X[i] / BX
8      k = i
9  Next
10 For s = 1 To k
11     If alfa[s] < alfa[s-1] Then
12         alfa[s] = alfa[s - 1]
13         X[s] = X[s - 1]
14         Descision[s] = Descision[s - 1]
15     End If
16     l = s
17 Next
18 return Descision[l]
19 return alfa[l]
    
```

Algorithm (2) Splitting The Information Table Into N Sequential Sub-modules

```
// Input: rough.table(U,A) is an information table
// Output: devide the main rough model into "N" modules

1   Read nofmodule
2   Define total, n, d, d2, case
3   total = rough.Tables(0).Rows.count
4   n = int(total / nofmodule)
5   d = total / nofmodule
6   d2 = (d - n) * nofmodule
7   For i = 1 To nofmodule
8       If i = nofmodule Then
9           n = n + d2
10          End If
11          For j = 1 To n
12              case = select min(cases) from rough.table(0)
13              R(case) = rough.Tables(0).Rows(case)
14              Rough.tables(i).rows(case) = R(case)
15              rough.Tables(0).Rows(case).clear
16          Next
17      Next
```

Algorithm (3) Splitting The Information Table Into N Random Sub-modules

```
// Input: rough.table(U,A) is an information table
// Output: devide the main rough model into "N" modules

1   Read nofmodule // number of modules
2   Define total, n, d, d2, case
3   total = rough.Tables(0).Rows.count
4   n = int(total / nofmodule)
5   d = total / nofmodule
6   d2 = (d - n) * nofmodule
7   For i = 1 To nofmodule
8       If i = nofmodule Then
9           n = n + d2
10          End If
11          For j = 1 To n
12              case = Int((total * Rnd()) + 1)
13              If Rough.tables(i).Rows(case).Rows.count > 0 Then
14                  GOTO 12
15              Else
16                  R(case) = rough.Tables(0).Rows(case)
17                  Rough.tables(i).Rows(case) = R(case)
18                  rough.Tables(0).Rows(case).clear
19              End if
20          Next
21      Next
```

Algorithm (4) Gating a decision from N Rough Decision modules

```
// Input: N modules of rough.table(U,A)
// Output: chosen decision, greatest  $\alpha$  degree

1  Read rule, nofdecisions, nofmodules
2  Define K, netdecision, netalfa, Descision[], C[], alfa[],
3  Define poss[],avgalfa[], avgposs[],decval[]
4  For I = 0 to nofdecisions
5  Input Descision[i]
6      For N = 1 to nofmodules
7          Run Algorithm 1. For rough.Tables(N)
8  If rough.Tables(N).getdecision(decision) = Descision[i] then
9  {
10      C[i] = C[i] + 1
11      Alfa[i] = alfa[i] + rough.Tables(N).getdecision(alfa)
12      Poss[i] = poss[i] + rough.Tables(N).getdecision(poss)
13  }
14  End if
15  Next
16      Avgalfa[i] = alfa[i]/C[i]
17      Avgposs[i] = poss[i]/C[i]
18      Decval[i] = Avgalfa[i] * Avgposs[i]
19      K = i
20  Next
21  For s = 1 To k
22      If decval[s] < decval[s-1] Then
23          Descision[s] = Descision[s - 1]
24      End If
25      l = s
26  Next
27  Netdecision = Descision[l]
28  Netalfa = decval[l]
29  return Netdecision
30  return netalfa
```

5. Real Case Study

As a case study used to explain the implementation of MRDM proposed model, we prepared a set of data obtained from a study of the impact of some factors on the vase life of carnation flowers including the following information factors collected by M. A. Fatah [12]:

- Effect of storage pretreatments
- Effect of silver thiosulfate (STS)
- Number of days from star stage to full opening stage (days).

Table 2 represents a part of the data collected for the 480 view cases for different samples representing the effects of the factors listed above. Table (3) shows the meaning of the used abbreviations of data items in table 2.

Table (2) Summary of information system for the given case study

StorTemp	UsingSTS	DFIOpen	VaseLf	number of cases
four	YES	less than 4	D	8
zero	NO	less than 4	D	24
four	NO	between 4 and 7	B	21
four	NO	between 4 and 7	C	35
four	NO	between 4 and 7	D	8
four	YES	between 4 and 7	B	24
four	YES	between 4 and 7	C	16
room	NO	between 4 and 7	A	10
room	NO	between 4 and 7	B	22
room	NO	between 4 and 7	C	17
room	YES	between 4 and 7	B	2
zero	NO	between 4 and 7	B	15
zero	NO	between 4 and 7	C	8
zero	YES	between 4 and 7	B	10
zero	YES	between 4 and 7	C	10
four	NO	between 7 and 9	D	8
four	YES	between 7 and 9	B	3
four	YES	between 7 and 9	C	22
room	NO	between 7 and 9	A	4
room	NO	between 7 and 9	B	12
room	NO	between 7 and 9	C	15
room	YES	between 7 and 9	A	15
room	YES	between 7 and 9	B	40
zero	NO	between 7 and 9	B	8
zero	NO	between 7 and 9	C	16
zero	YES	between 7 and 9	A	1
zero	YES	between 7 and 9	B	19
zero	YES	between 7 and 9	C	14
four	NO	between 9 and 12	C	8
four	YES	between 9 and 12	C	7
room	YES	between 9 and 12	B	23
zero	NO	between 9 and 12	C	9
zero	YES	between 9 and 12	B	9
zero	YES	between 9 and 12	C	17
Number Of Cases				480

Table 3. Data Items Description of the information system

Abbreviation	Description	Data Type
StorTemp	Effect of storage temperature	Accept one of the following data: – Room temperature – Zero ° C – four ° C
UsingSTS	Using silver thiosulfate (STS)	Accept Yes or NO
DFIOpen	Number of days from star stage to full opening stage	This stage lasts from one day to 12 days and this period may be divided into categories as follows: – less than 4 – between 4 and 7 – between 7 and 9 – between 9 and 12

VaseLf	Vase life, it is a decision attribute and its mean is life of the flower after cut	Accept value of this data set: {A, B, C, D} Vase life period classified into four categories D means that vase life less than 10 days C means that vase life is greater than 10 days but less than 20 days B means that vase life is greater than 20 days but less than 30 days A means that vase life over than 30 days
--------	--	--

The steps of algorithm (1) proceed as follows:

- Step 1: define inference rules which are given to take a decision.
 Step 2: which is done by (MRDM) proposed model is determining upper and lower approximation for all possible cases to choose the optimum decision.
 Step 3: calculating α degree for the different decisions according to the given rules.
 Step 4: chose the optimum decision which has the greatest α degree.

In our example, let the given rules is:

StorTemp = 'room' and DFIOpen ≥ 9

Possible decisions VaseLf = {'A', 'B', 'C', 'D'}

B = {StorTemp, DFIOpen}

X = {x | if StorTemp = 'room' and DFIOpen ≥ 9 then VaseLf = 'B'}

$P|BX| = 23$, $P|X| = 23$

$$\alpha = \frac{P|X|}{P|BX|} \quad \alpha = \frac{23}{23} \quad \alpha = 1 \quad (1)$$

Y = {x | if StorTemp = 'room' and DFIOpen ≥ 9 then VaseLf = 'A'}

Z = {x | if StorTemp = 'room' and DFIOpen ≥ 9 then VaseLf = 'C'}

K = {x | if StorTemp = 'room' and DFIOpen ≥ 9 then VaseLf = 'D'}

$P|Y| = P|Z| = P|K| = 0$

$$\alpha = \frac{P|Y|}{P|BX|} = \frac{P|Z|}{P|BX|} = \frac{P|K|}{P|BX|} \quad \alpha = 0 \quad (2)$$

From 1, 2 $\alpha(X)$ is the greatest

The optimum decision is VaseLf = 'B' with $\alpha = 1$

The steps of algorithm (2) proceed as follows:

- Step 1: determining the number of module needed to create from the main module, the user of MRDM do this step.
 Step 2: calculate number of cases in each module (N cases)
 Step 3: select the first N cases from the main information system and insert them into module 1, this step is repeated as the number of modules determined in previous step.

The steps of algorithm (3) proceed as follows:

- Step 1: determining the number of module needed to create from the main module, the user of MRDM do this step.
 Step 2: calculate number of cases in each module (N cases)
 Step 3: select one case randomly from the main information system and insert it into module 1, this step is repeated until it is full of the module.
 Step 4: the previous step is repeated as the number of modules determined in step 1.

After creating modules from the main rough information system, MRDM proposed model allows user to define rules, which are needed to get a decision. Note that the same example used in taking decision from main rough information system represented in table (2), is used to represent using MRDM proposed model to take a decision through modularity, and also given rules are used. In our example, the main information system has been converted to four models using the random mechanism. Table (4) represents summary data of the four modules, and our given rule is the same rule used above.

StorTemp = 'room' and DFIOpen >= 9.

Table 4. Summary of Rough modules data

StorTemp	UsingSTS	DFIOpen	VaseLf	number of cases				
				M1	M2	M3	M4	Main
four	YES	less than 4	D	5	0	2	1	8
zero	NO	less than 4	D	5	10	5	4	24
four	NO	between 4 and 7	B	2	5	8	6	21
four	NO	between 4 and 7	C	7	9	6	13	35
four	NO	between 4 and 7	D	1	2	3	2	8
four	YES	between 4 and 7	B	6	5	6	7	24
four	YES	between 4 and 7	C	0	6	2	8	16
room	NO	between 4 and 7	A	3	5	2	0	10
room	NO	between 4 and 7	B	7	5	7	3	22
room	NO	between 4 and 7	C	3	4	7	3	17
room	YES	between 4 and 7	B	0	0	1	1	2
zero	NO	between 4 and 7	B	4	4	2	5	15
zero	NO	between 4 and 7	C	3	2	2	1	8
zero	YES	between 4 and 7	B	4	4	1	1	10
zero	YES	between 4 and 7	C	3	2	1	4	10
four	NO	between 7 and 9	D	1	3	3	1	8
four	YES	between 7 and 9	B	2	1	0	0	3
four	YES	between 7 and 9	C	3	6	8	5	22
room	NO	between 7 and 9	A	1	1	2	0	4
room	NO	between 7 and 9	B	5	3	1	3	12
room	NO	between 7 and 9	C	4	2	6	3	15
room	YES	between 7 and 9	A	3	4	3	5	15
room	YES	between 7 and 9	B	11	8	12	9	40
zero	NO	between 7 and 9	B	1	2	3	2	8
zero	NO	between 7 and 9	C	4	1	4	7	16
zero	YES	between 7 and 9	A	0	0	1	0	1
zero	YES	between 7 and 9	B	5	6	6	2	19
zero	YES	between 7 and 9	C	4	2	3	5	14
four	NO	between 9 and 12	C	2	3	1	2	8
four	YES	between 9 and 12	C	2	1	2	2	7
room	YES	between 9 and 12	B	8	2	5	8	23
zero	NO	between 9 and 12	C	2	2	3	2	9
zero	YES	between 9 and 12	B	4	4	0	1	9
zero	YES	between 9 and 12	C	5	6	2	4	17
Number Of Cases				120	120	120	120	480

After defining rules, one decision is taken from each module with α degree. Final step in taking decision through MRDM proposed model using modularity approach is gating process. In MRDM proposed model voting technique is used as a gating process, this done by making vote between the decisions taken by the modules. The voting process is taking into account two factors which are α degree and possibility degree. Possibility degree is calculated in MRDM proposed model as percentage between the numbers of cases achieve the given rule to number of cases achieve chosen decision

The steps of algorithm (4) proceed as follows:

Step 1: Using algorithm (1) to get a decision from each module

Step 2: Calculate α degree and possibility degree for the optimum decision of each module as following:

$$\alpha = \frac{P | X |}{P | \bar{B} X |} \text{ where } X = \{x \mid \text{number of cases achieve the optimum decision according to the given rule}\} \text{ and } \bar{B} X \text{ is the upper approximation of the given rules}$$

$$\text{Possibility} = \frac{P | X |}{P | D |} \text{ where } D = \{x \mid \text{number of cases achieve the optimum decision}\}$$

Step 3: apply voting technique to select the optimum decision of the model among the decisions achieved in each module, according to α degree and Possibility degree

The vote of each decision is calculated as a summation of [α * possibility] for each module satisfies the decision, figure (3) explains how to implement the voting process.

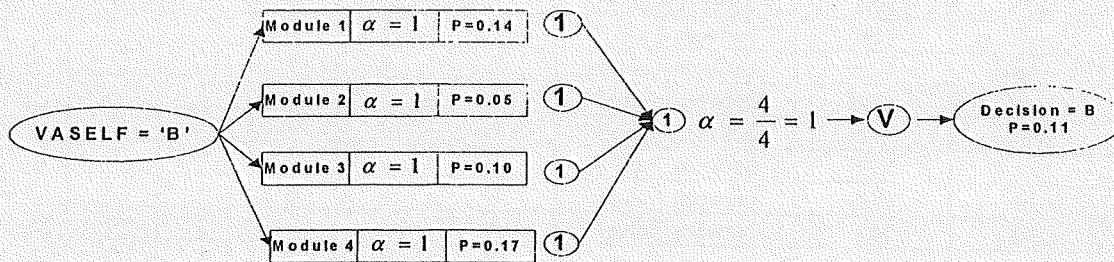


Figure. 3 voting process implementation

6. CONCLUSION

Modular approach reduces the computation cost of drawing decision of large rough decision models. A gating technique can be proposed by the user to aggregate the output decisions of the various decision modules so as to get the overall output decision. In this paper we adopt a voting technique as the gating method. The illustrative example shows that the design effort and computation needed in each subsystem design are typically lower. Also, the system will be easier to debug and maintain. As a future work.

As a future work, we plan to propose a "user modular: approach of the above MRDM through which fuzzy terms are used to replace the continuous real values in the given information system. This will result in a further reduction of the computation cost for drawing the final decision output of the rough decision model.

REFERENCES

- 1- C. Degang, Z. Suyun, "Local reduction of decision system with fuzzy rough sets", *Fuzzy Sets and Systems* 161 (2010) 1871–1883
- 2- H. C. Tseng, B. Almogahed, "Modular neural networks with applications to pattern profiling problems", *Neurocomputing*, 2008.10.020.
- 3- P. Melin, C. Gonzalez, D. Bravo, F. Gonzalez and G. Martinez, "Modular Neural Networks and Fuzzy Sugeno Integral for Pattern Recognition", *Studies in Fuzziness and Soft Computing*, Vol 208 (2007), 311-326.
- 4- R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts", *Neural Computation*, 3:79:87, 1991.
- 5- M. I. Jordan, R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, 6:181-214,1994.
- 6- Z. Pawlak, "Rough sets", *International Journal of Computer and Information Sciences* 11 (1982) 341–356.
- 7- Z. Pawlak, A. Skowron, Rudiments of rough sets, *Information Sciences*, *Information Sciences* 177 (2007) 3–27.
- 8- K. Thangavel , A. Pethalakshmi, " Dimensionality reduction based on rough set theory ", *Applied Soft Computing* 9 (2009) 1–12
- 9- Y. Qian , J. Liang, D. Li , H. Zhang, C. Dang, "Measures for evaluating the decision performance of a decision table in rough set theory", *Information Sciences* 178 (2008) 181–202.
- 10- J. W. Grzymala-Busse, S. Siddhaye, " Rough Set Approaches to Rule Induction from Incomplete Data", the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4–9, 2004, vol. 2, 923–930
- 11- Y. Qian , J. Liang, D. Li , F. Wang, N. Ma, "Approximation reduction in inconsistent incomplete decision tables", *Knowledge-Based Systems* 23 (2010) 427-433.
- 12- M. A. Fatah, "Effect of some postharvest treatments on carnation flowers quality", Ph.D. Thesis Agric. Sci. (Ornamental Horticulture), Cairo university, (2011).

THE UNIVERSITY OF CHICAGO
LIBRARY
540 EAST 57TH STREET
CHICAGO, ILL. 60637



A Flexible Data Integration Framework based on Semantic Web Engineering

Ibrahim M El- Henawy¹, Mahmoud M. Abd Ellatif² and Tamer E. Amer²

Abstract:

The huge amount of data is a common in electronic environment now. Usually, these Data are distributed over different sources and formats. So data management is faced by problems of integrating data, and extracting relevant information. Traditional data integration approaches have many obstacles and the semantic web engineering can help in solving many problems of data integration. Therefore, this paper displays an integrated framework to simplify the process of data integration. The proposed framework consists of four layers: data source, mapping, semantic, and application. The paper shows different web technologies that are used to build framework parts.

Keywords: *Semantic Web, Resource Description Framework (RDF), RDF schema, Ontology, Data Integration*

1 Introduction

Today's business environment is characterized by the challenges of increased globalization. Therefore, the widespread characteristic among profit or nonprofit organizations is a collaborative to achieve their goals. For collaboration to be successful, the need to share information has increased. Therefore, the need to interorganizational information system (IOS) and Collaborative Information Environment (CIE) have become more important. Therefore, using IOS and CIE have increased continuously, because it represents a backbone for any collaboration projects. Enterprises use IOS to implement banks networks, e-government, financial markets, Supply Chain Management, and Customer Relationship Management.

One of the critical success factors for IOS is high integration among internal information systems. However, enterprises have heterogeneous data sources, so the main problem is how to integrate diverse data sources. An important step is to formulate an integration map, which specifies how to extract, transform, and load the data from the local data sources into a global data store.

The integration of information systems are formed in four phases [1]:

1. Study and analyze the data schema of local data sources in detail, to determine the semantics of the data.

¹ Faculty of computer and Information, Zagazig University, Zagazig, Egypt

² Faculty of Computer and Information, Mansoura University, Egypt

2. Convert the heterogeneous representations of local data (e.g. Entity-Relationship-Models or XSchemata) into a common global data model to overcome conflicts resulting from varying model concepts.
3. Make *integration* schema through the correspondences among the elements of the local schemas, and integrating these source schemas into a global; unified schema.
4. Finally, the results of the pervious phase are used to consolidate the local stored data in a way; therefore, the integrated data can be accessible based on the global schema of data (integrated schema).

In conclusion, data integration is the process that gives users access to multiple heterogeneous data sources through queries on a unified schema. It involves combining data that are stored in different resources and providing a unified view of these data for users [2]. Figure 1, represents data integration process.

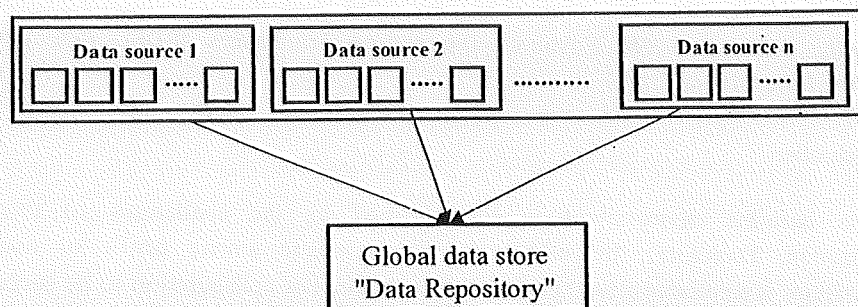


Figure (1): Data integration process

The most important issue in data integration is the Schema integration. How can data analyst make match up among equivalent real-world entitics from multiple data sources? This is referred to entity identification process. Terms maybe give different interpretations at different sources. For example, data analyst will find *customer_id* and *cust_number* in different two databases are the same attribute [3].

IOS and CIE will depend on our ability in developing good solutions for the problems of data integration. The solutions of data integration problems should be transparent to the user; therefore, these solutions should mask far from the user, and implement on the underlying data sources. Ideally, they make the set of different data sources look like a single data source to the user. Data integration requires combining and matching information in different sources, and resolving a variety of conflicts. The user does not aware where the data is stored, or what language of programming interfaces that is supported by the system of data source [4].

Semantic web engineering offers a framework and technologies to share, manage, and reuse data across applications, enterprise, and community boundaries [5]. Semantic Web

technologies can be used in data integration. Various locations and different formats Data can be integrated to use in many fields such as [6,23]:

- Resource discovery and classification to improve the capabilities of search engine in specific domain;
- Cataloging for describing the content and content relationships available at a particular Web site, page, or digital library;
- Knowledge management to facilitate knowledge sharing and to exchange by intelligent software agents ;
- Content Management: for describing collections of pages, which represent a single logical “document”, and intellectual property rights of Web pages.

The paper outline is as follows: section 1 provides an overview about data integration process. The traditional data integration techniques and the challenges for applying these techniques are discussed in section 2. Section 3 illustrates what is the semantic web and its core technologies such as Resource Description Framework (RDF) and Ontology. In section 4 the semantic data integration framework shows the functions of its main components. A discussion is made about the framework in section 5. Paper conclusion and the future work are presented in section 6.

2 Traditional Data Integration Approaches

In general, data integration combines the data from different sources and provides a global view in unified schema to service the users. The wide range of available data sources need to integration techniques to integrate data efficiently.

Through Literature review, some authors classified data integration approaches into four categories; integration by application, integration by middleware, uniform data access, and common data storage [4,7, 24]. Each approach has several strengths and weaknesses, which show its applicability to solve the problem of data integration. Authors in previous works suggested several Data Integration Mechanisms based on data integration approaches. These Mechanisms include : data warehousing , federation database systems, Operational Data Stores (ODS), Message Oriented Middleware (MOM), Service-Oriented-Architecture (SOA), Enterprise Service Bus (ESB), Object-Oriented Databases, and Peer-Data-Management Systems[24].This section introduces two famous traditional data integration Mechanisms, which are: data warehousing and federation database systems.

2.1 Data Warehousing

Data warehouse (DW) system extracts, cleans, conforms, and delivers source data into a dimensional data store, to implement querying and analysis. Data warehousing is

transforming data from legacy and transaction database systems into organized information, in a user-friendly format. Data warehouse (DW) system encourages data analysis and support fact-based business decision-making. The process of transforming data from its original format to a dimensional data store consumes at least 70 percent of time, effort, and expense of data warehouse projects [3].

Data warehouse designers, based on business needs, define a global schema for the shared data. They provide the logic to restore data and pump it into a centralized repository (the data warehouse). An extract-transform-load (ETL) process extracts the data from the identified original databases, transforms disparate formats into a single format, and loads the results into the data warehouse. The user accesses to the data warehouse to extract information. Typically, the user can read data warehouse only, but updates make directly on the source systems. Sometimes data marts give individual communities their own subsets of the global data [26].

2.2 Federated Database

Federated database systems acts as a virtual data warehouse system. But the main difference between them is the federated database systems retain the physical data and a middleware layer (adaptors) translates all requests to run against the source systems. The adaptors assume that they control over the query process. However, query performance and index types are mainly under control of the data sources. Adaptors are usually custom-designed for every data source and involve a "one size fits all" front-end query data schema. Access controls can be put in place. The adaptors convert general queries to the correct syntax for the individual databases, individual database queries are executed, and individual database results are processed and presented in a universal format.

Typically, a data source has existing applications and users. Therefore, it is important the operation of the source did not affect when it is brought into a federation. Existing applications should run unchanged, data should be neither moved nor modified, interfaces should remain the same, and no additional software should be installed on the machine that hosts the data sources. The way of data source processes requests for data should not be affected by the execution of global queries against the federated system, though those global queries may be in touch with many different data sources. Likewise, there should be no impact on the consistency of the local system when a data source enter or leave a federation. Communication with the data source is via a client-server architecture using the source's normal client [4].

2.3 Obstacles

Traditional data integration approaches, including data warehousing and data federation, involve mapping between data models, but they have many obstacles. Systems designers

prefer Data warehouses than federated database often. Because Data warehouses can support several operations, include clean and transformed (usable) data, a wide range of queries, fuzzy matching, Name variation, nearest neighbor, Complex queries without shutting down data source systems, and aggregations and other calculations. But Federated database cannot support any of these operations, plus fetching data for applications and populating associated databases.[7,25]

In another side, data warehousing needs to Extraction, Transformation, and Loading (ETL) process to keep a copy of data from all underlying data sources in a synchronized. Therefore data warehouse faces difficult in implementation, expensive to keep data, and it is inflexible to change in the queries that asserted against the data. Any data not available in data warehouse not conveniently queried. Data that are extracted for the warehouse needs to remodel. Some places that have large upfront design burden on data warehouse schema and the ETL process. In addition, Sarbanes-Oxley law (USA) obliges companies to store and access original data or copies of original data, not the modified data that typically resides in data warehouses [25]. Fundamentally, data warehouses are designed to answer set of specific questions repeatedly as new data become available. Data warehouses do not themselves contain any semantic knowledge. For example, the responsibility of knowing, which two compounds with different names are structurally related, is on the user composing of the questions [7].

Traditional data integration Mechanisms face significant challenges such as varying levels of granularity , highly changeable in data sets, incompleteness and noisy of data [7]. If one of the underlying databases changes, the integrated database needs to update. In addition, the unified schema needs to modify if data model of any underlying databases was updated, so the integration process must be an ongoing task [8]. In fact, data warehouses seem to be in conflict with global 24*7, real-time interactions and the Service-oriented architectures (SOA) [25].

There are also technical challenges; for example, in most of domains there are different databases in different schemas that are serving different applications [9]. The external views, which are presented for the user by each DBMS, are different according to the schema that DBMS utilize, while traditional approaches do not provide a unified schema for data access.

3 Semantic Web Engineering

The main goal of semantic web is publishing data in a form can be processed by the machine. Semantic web, a new generation of web, is an extension of the current web in which information includes well-defined meaning. Semantic web enables computers and people to work in cooperation effectively [10]. There are many technologies to build

semantic web solutions, this section, briefly, describes most important technologies that are used in the presented framework.

3.1 Resource Description Framework (RDF)

RDF is a data model for providing metadata about web such as title, author, modification date, and copyrights. It represents the core technology of the semantic web applications [11, 12, and 13].

The goal of RDF is overcome the problem related to lack of semantic representation in XML to facilitate interoperability among applications. RDF provides abstract and common conceptual framework, to express information in a way accessible by machines. Therefore, RDF joins various applications in distributed setting without loss of meaning [14, 15]. it could be available to apply other than original forms of data, and it becomes a World Wide Web Consortium (W3C) standard recommendation in 2004[16].

The RDF model consists of two main parts; RDF and RDF schema. RDF is the language for connecting metadata with the document parts; while, RDF schema is a terminology description language for RDF that defines the possible vocabularies and the grammar for how can form these metadata? Web designer uses RDF to create ontology of specific domain.

3.2 Ontologies

Ontology is the heart of semantic web applications. Defining ontology [15] is “the explicit and formal specification of conceptualization of a domain of interest”. It represents a key technology for enabling semantics-driven knowledge processing. Communities establish ontologies, or share conceptual models, to provide a framework for sharing a precise meaning of symbols exchanged during communication and enable software to give reason about different worlds and environments [18, 19].

Any Ontology contains Classes to define set of concepts; these classes have a set of characteristics and attributes (properties). The main goal of ontology engineering is trying to capture the semantics of a particular domain of knowledge that reflect the view of the human, and putting it in the readable form by machine. This process means that the computer can interpret the semantics within the ontology without direct human involvement. It also characterizes the semantics in terms of concepts and relationships that relate these concepts to each other [20].

The notion that communities are establishing ontologies to share conceptual models [19]. They will make easy communication among them based on their defined ontology. This

notion of the sharing knowledge among communities will emphasize the fact that people have a common agreement with the meaning which expressed by the ontology; hence the importance of agreement appears.

4 Semantic Data Integration Framework

The framework presented by the proposed solution can be considered as integrated semantic web framework. The word “integrated” refers to the architecture consists of more than one component, each of which has a specific task. That is “semantic” means that the architecture is based on semantic web technologies to perform the semantic integration process [21].

The proposed framework approach goes beyond how the data is modeled in its local source to carry meanings. It uses the semantic web technologies to build ontology, which concerned about the semantics being conveyed to integrate different data source. One of the fundamental advantages of ontologies is their relative independence from applications. Ontologies can be reused by different kind of applications. Several ontology languages based on XML can be used to encode ontologies such as OWL [22]. The developed ontology provides a machine-readable representation about data resident in different resources for the application that performs data integration.

As shown in Figure 2, this framework is based on a layered approach. It consists of four layers, which are summarized as follows:

- **Data source layer:**

This layer represents different resources that provide the data to the semantic web application. Data resources could be: databases with similar or different schemas; spreadsheets (.xls) files; text files (.doc) files; images (scanned brochures); Portable Document Format (PDF); or data in semantic web formats such as RDF Site Summary (RSS).

- **Mapping Layer:**

This layer aims to transform the data from its original format, in local resources, to the semantic web format in RDF or OWL. It contains three basic components:

- **Relational Database to semantic web mapping (RDB2Semantic):** this component aims to present the relational database resources to the semantic web. An interesting tool called D2RQ can do this step.
- **Annotation module:** this module aims to encapsulate the resources in the form of text files, PDF files, or images with semantic web data using the Dublin Core Metadata (DCM).
- Resources in RDF format do not need to transform in the mapping layer, so it will go through the next layer without any action.

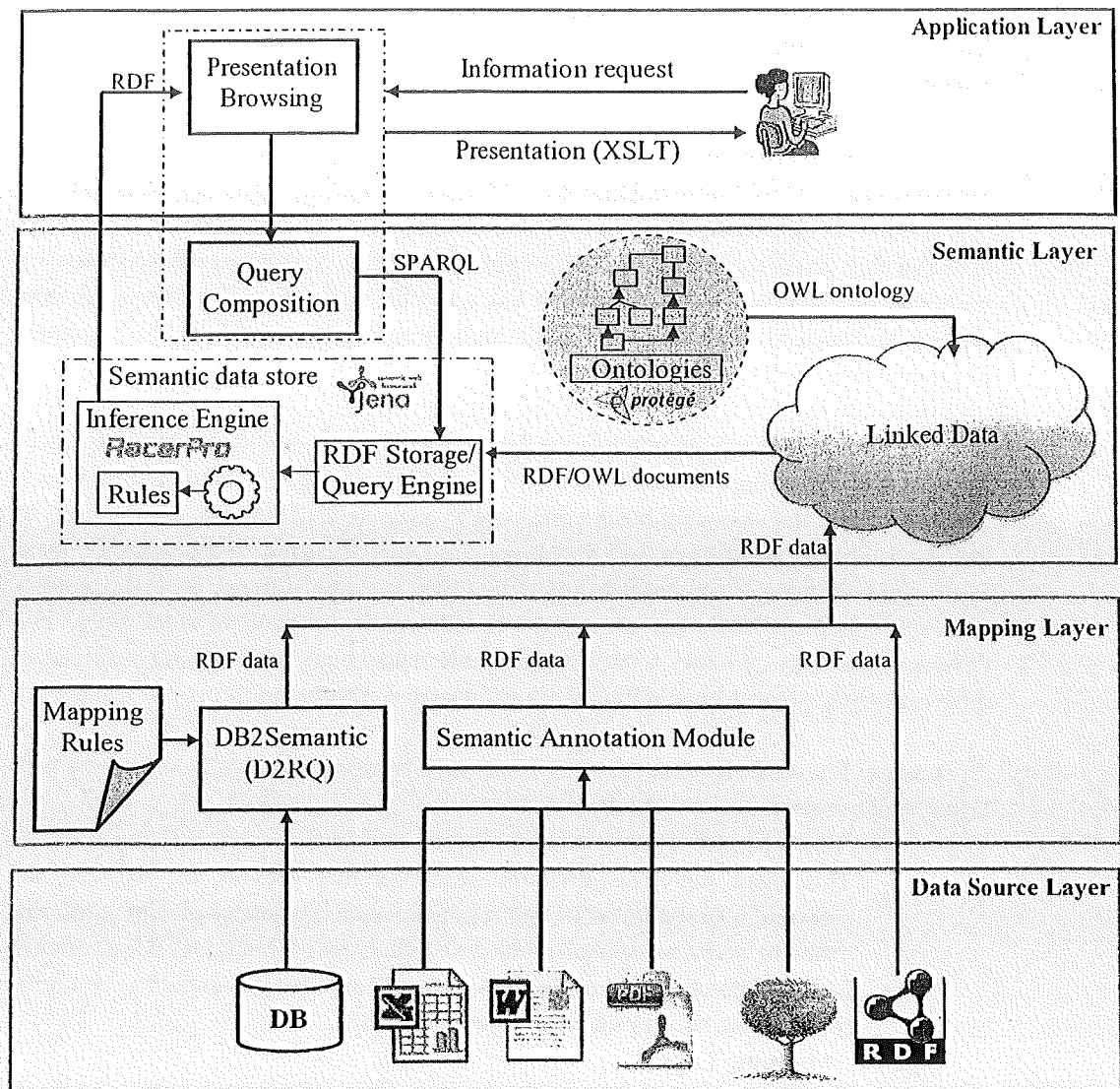


Figure (2): Semantic web framework

- **Semantic Layer:**

This layer holds the semantic web data that is resulted from the mapping layer. It contains different components as follows:

- **Linked data Cloud:** it contains the data represented in the semantic web standards such as RDF and OWL whether it is transformed via set of modules or tools in the mapping layer or it

comes from the data source layer in the form of linked data; such as data comes from RDF Site Summary (RSS) channels.

- **Ontologies:** It is ontologies structures that are built via ontology, editing environment such as protégé. These ontologies hold the vocabularies and terms by which the data stored in different resources are described.
- **Semantic data Store:** This component plays the role of semantic repository that holds the ontology model. It consists of two main parts: the first part is for storage and query engine; to manage the queries against the stored RDF or OWL documents. The second part is the inference engine; to make reasoning against the stored ontologies based on specific rules. The racer is used in this part. The semantic data store component can be represented by Jena Framework to store and manage ontologies developed in the system.

- **Application Layer:**

This layer is responsible for representing the application logic and the graphical user interface. It is used to send required information against the stored linked data, and to receive presentation in the form of HTML markup. It is noticed that the Graphical User Interface (GUI) component spans in the application layer and the semantic layer. This component consists of two parts: the first part is the presentation and browsing which responsible for interacting with the application users. This part will be in the application layer. Query composition is the second part. It transforms user's query from the human readable form to another that could be run on the stored linked data, and returns results to the application users. The query that is run on the stored linked data will be in the SPARQL format. SPARQL is a query language for RDF. For this reason, this part is resident in the semantic web layer. Jena framework also supports querying facilities. The output of the query will be in the form of RDF/XML format, so the role of the presentation and browsing part is to uses a technology such as XSLT to provide the human readable form of the queried data.

5. Discussion

So far, this paper has not introduced the framework, which is built up on semantic web technologies to facilitate data integration process in a certain domain only. But also, the paper discusses the technologies and tools that are used to achieve this framework. The decision of using the semantic web to integrate data in a specific domain is better than

design new web sites from scratch. The existing systems will be adapted to interact with the semantic web technologies. Software developers will make the following steps to implement the proposed semantic data integration solution:

1. For data that are extracted from relational database:
 - Use protégé, as one of the most famous ontology editing, to develop the target ontology for each database resource according to the structure of the relational database.
 - Use the D2RQ server to map automatically the relational database to the corresponding target RDF or OWL ontology structure. In this step, software developers specify the mapping rules according to the original database schema and their corresponding ontology structure.
 - Integrate the resulting linked data into a single shared ontology and semantically link the ontological terms using OWL.
2. For the resources that publish its data as spreadsheets, Word files, PDF, or images, run the semantic annotation module to envelop these resources with the semantic metadata such as Dublin Core Metadata.
3. For the resources that publish its data as linked data in RDF format such as RSS channels, extract the linked data as it is. it does not need any transformation
4. Use Jena to store, manage the generated ontologies, and linked data.
5. Search capabilities using SPARQL RDF query Language and reasoning can be done within Jena and Racer reasoned.
6. Use XSLT to transform the machine readable form of the query result to the human readable form in HTML Format

Table 1 summarizes the technologies and tools used in this framework as follows:

Table 1: Tools and technologies used in the framework

Technology/tool	Purpose
RDF/OWL	languages for representing interchanging data
D2RQ server	to present the database systems to semantic web
Protégé	an editing environment to build the ontologies used in the semantic web application project
Jena	Open source framework for persistence storage for the ontologies.
SPARQL	RDF query Language to build queries against ontologies
XSLT	Transform the query results to human readable data

Table 2 illustrates the comparison between traditional data integration Mechanisms and proposed framework. The properties of data integration mechanisms are divided into two categories. The first deals with properties related to the reliability of a data integration mechanism, and the second is its flexibility. Reliability refers to the capability of a data

integration mechanism to perform its function correctly in a specified period under stated operation conditions. Whereas, Flexibility refers to the ability of a data integration mechanism to, react flexibly to change in the surrounding system environment, and be applicable to disparate types of data source. [24]

Table 2: Compare between traditional and semantic approaches

Criteria	Metrics	Data Warehouse	Federated Database	Semantic based Proposed Framework
Reliability	Timeliness(Ability to guarantee up-to-date data)	√	√√√√	√√√√
	Reactivity(Ability to guarantee a timely response to a query)	√√√√	√√√√	√√√√
	Robustness (Ability to function reliably under any circumstance)	√√√√	√√√	√√√√
	Quality of data	√√√√	√√√√	√√√√
Flexibility	Data volume scalability (Ability to manage increases in data volume)	√√√√	√√√	√√√√
	Data source scalability (Ability to manage increases in the amount of data sources)	√√√√	√√	√√√√
	Data source agnosticism (Ability to integrate different types of data source)	√√√	√√√	√√√√
	Adaptability (Ability to react to changes in data sources)	√√	√√	√√√√
√ Weakly	√√ Slightly	√√√ Averagely	√√√√ Largely	√√√√ Strongly

6 Conclusions and Future Work

This paper presented a semantic web based framework for data integration in a certain domain. It stated the main challenges in the traditional data integration techniques such as data warehousing and federated databases during implementation. The paper aimed to address this problem and proposed a data integration framework based on semantic web technologies to solve it. The main idea in this solution was to enable the data in different resources to be annotated using standard vocabularies of a shared ontology. This created a state of agreement among different resources; all of them dealt with the different data categories via unified agreement structure. Briefly, the most important advantages of proposed framework were no limited to integrate data sources, and treatment the different forms of data.

The presented framework consists of four layers in which each layer in the upper level based on a layer in the lower level; this of course will reduce the design complexity. Data source layer contained the different data categories used by the application. Mapping

layer transformed the data from data source layer into the data in the form of linked data. Semantic layer contained all the semantic web data in the form of RDF or OWL files. Finally the application layer which responsible for application logic and GUI. The framework showed different semantic web technologies used to build its component.

We suggest two directions to expand the presented work through focus on:

- How can web developers convert data in the data source layer into the semantic web?, and show different techniques that are used to make this mapping,
- The role of ontology: to illustrate how can web developers build the shared ontology as a base for data integration process.

References

- [1] Stefan kurz, et.al., "A UML profile for modeling schema mapping", Institute for Information Systems and Software Technology (IFIS) University of Passau; German, www.fim.uni-passau.de/fileadmin/files/.../KGF06.pdf, 2006
- [2] Ladjel Bellatreche, Guy Pierra, Dung Nguyen Xuan, Dehainsala Hondjack and Yamine Ait-Ameur, "A Priori Approach for Automatic Integration of Heterogeneous and Autonomous Databases", International Conference on Database and Expert Systems, 2004.
- [3] Kalinka Mihaylova Kaloyanova, " Improving Data Integration For Data Warehouse: A Data Mining Approach", Faculty of Mathematics and Informatics, Sofia University of Sofia, Bulgaria 2006. www.nbu.bg/PUBLIC/IMAGES/File/.../17.pdf
- [4] Kevin Brandt, Paul Fernan, Christian Grant and Russell Richards, " Operational Net Assessment Enabling Technology: Data Integration ", Project Alpha technical report. 2004.
- [5] C. Pasquier , "Biological data integration using Semantic Web technologies", Biochimie 90,2008.
- [6] D. Dou, Han Qin, and Paea Lependu, " ONTOGRATE: Towards Automatic Integration for Relational Databases and the Semantic Web through An Ontology-Based Framework", International Journal of Semantic Computing , Vol. 4, No. 1 ,2010.
- [7] Stephen P.Gardner, " Ontologies and semantic data integration", DDT.,Volume 10, Number 14, 2005.
- [8] Lincoln D. Stein, "Integrating biological databases", Nature Publishing group volume 4, 2003.

- [9] Ibrahim M El-Henawy, Mahmoud Abdellatif and Tamer E. Amer, "Modeling the Egyptian Tax Authority using RDF", In Proceedings of the the 45th annual conference in statistics, operation research and computer science. Institute of Statistical Studies and Research, Cairo University, pp.106-118, 2010.
- [10] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, Vol. 284, ,PP.28-37, 2001.
- [11] Frank Manola and Eric Miller, "RDF primer", W3C Recommendation, W3C 2004.
<http://www.w3.org/TR/REC-rdf-syntax/>
- [12] Ora Lassila and Ralph R. Swick. "Resource description framework (RDF) model and syntax specification", Technical report, W3C Recommendation, 1999.
<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>
- [13] Graham Klyne and Jeremy J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax". W3C Recommendation, 2004.
<http://www.w3.org/TR/rdf-concepts/>
- [14] Steffen Staab, Michael Erdmann, Alexander Maedche, and Stefan Decker, "An Extensible Approach for Modeling Ontologies in RDF(S)", AIFB, Univ. Karlsruhe, Germany, 2000.
http://www.ics.forth.gr/isl/SemWeb/proceedings/session2-1/html_version/index.html
- [15] Mohammad Mustafa Taye, "The State of The Art: Ontology Web based Languages: XML based", Journal of Computing, Volume 2, Issue 6, ISSN 2151-9617. June 2010.
- [16] Amit Sheth, "Semantic Web: Technologies and Applications for the Real-World", Kno.e.sis Center, Wright State University, WWW conference, Banff, Alberta, Canada, May 2007
- [17] T.R. Gruber, "Towards Principle for the design of ontology used for knowledge sharing", International Journal Human-Computer Studies 43, pp.907-928, 1993.
- [18] Edd Dumbil, "Building the Semantic Web", O'Reilly media, 2001.
<http://www.xml.com/pub/a/2001/03/07/buildingsw.html>
- [19] Simone A. Ludwig and S.M.S. Reyhani, "Semantic approach to service discovery in a Grid environment", Journal of web semantics, Volume 4 Issue 1, January, 2006.
- [20] Michael C. Daconta, Leo J. Obrst and Kevin T. Smith, "The Semantic Web - A Guide to the Future of XML, Web Services, and Knowledge Management", Wiley publishing, Indiana, USA, 2003.
- [21] Ibrahim M ElHenawy, Mahmoud M. Abdelatif, Tamer F. Amer, "A Proposed Ontology Based Architecture to enrich the data semantics syndicated by RSS techniques

in Egyptian Tax Authority", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 9, PP. 203-209, 2010.

[22] Natalya F. Noy and Deborah L. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

<http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html>

[23] L. Chen, N.R. Shadbolt and C.A. Goble, "A Semantic Web Based Approach to Knowledge Management for Grid Applications", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 19, No. 2, 2006.

[24] Karl A. Hribernik; Christoph Kramer; Carl Hans and Klaus-Dieter Thoben, "Impacts of data integration approaches on the limitations of autonomous cooperating logistics processes", <http://www.econis.eu/PPNSET?PPN=663235138>, 2011.

[25] Gavin Robertson, "Dataless hybrid approach solves many universal data integration, sharing and interoperability problems", *Information Management Magazine*, formerly *DM Review*, March 2007.

<http://www.information-management.com/infodirect/20070316/1077688-1.html>

[26] Stefan B. , Wikan D. and Sunindyo T. M., Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects , 2010 International Conference on Complex, Intelligent and Software Intensive Systems, pp.360-367, IEEE 2010.

Enhancing Text Information Extraction in Videos Using WordNet library (AFS)

Sherif F.younis¹

Fatma A. El-Licy²

Akram I. Salah³

Abstract

The text data present in images and videos contains certain useful information for indexing and automatic annotation. This paper offers a new dimension to the frameworks for the extraction of text from video. Employing the new component WordNet library facilitates the process of extracting semantic from complementary resource. This paper integrates a new text extraction approach with the Wordnet Library to generate a metadata about the video content. AFS is a framework for text information extraction in video. The framework architecture composed of a standard Text Information Extraction (TIE) component, a general ontology component and a post-processing module. The TIE represents a primary complementary resource (pcr). Whereas, WordNet component is employed to represents general ontology as a second complementary resource (scr). The post-processing module integrates these two components in order to compile the extracted information supported by the ontology. The result of post-processing is a representation of the extracted information as metadata in a standard machine readable format. Use *AFS* facilitates better search for videos by any search engine.

Keywords: Primary complementary resources, Complementary resources, WordNet, Ontology, image processing, metadata.

1. Introduction

With the growing number of digital multimedia libraries, the need for efficient indexing, browsing and retrieving multimedia information have increased. Several approaches have been developed for indexing, querying and retrieving multimedia information. To facilitate these processes meta data and annotated data are needed

¹Institute of Statistical Studies and Research Computer Science Department Cairo University

to make referencing and searching Easier The textual information embedded in the multimedia data offers important information for multimedia data understanding. Therefore, it could be a good entity for keyword-based queries.

Text extraction has gained a lot of attention, in the last years, in several applications. Jung et al. [1] distinguish between four main classes of applications as follows:

- a) page segmentation (which include newspaper headlines extraction [2]);
- b) Address block location ;
- c) License plate location [3]; and
- d) Content-based image/video indexing [4].

In spite of Jung's efforts, the research in the field of designing domain independent text extraction systems is still a great challenge. The challenge relies on the diversities of the written text in the video sources. The written text could be in different fonts, size, color, orientation and alignment. Also, it could, possibly, be embedded in shaded or complex backgrounds.

1.1 Artificial and Scene Text

Text appearing in images and videos are, usually, categorized either as artificial or scene text, according to the its original source.

- Artificial text: It is referred to as caption text or superimposed text. As shown in Fig. 1, the text was not part of the image, but it is laid over the image at a later stage (e.g., the name of someone during an interview, the name of the displayed place, a notice or a broadcasting , etc.)



Figure 1: Artificial text

- Scene text: it is referred to also as graphics text. In contrast to artificial text, scene text is part of the scene, as shown in Fig. 2, (e.g., the name of a product during a commercial break). Scene text is, usually, more difficult to be extracted due to different alignment, size, font type, light condition, complexity of the background and distortion. Moreover, the text in the scene is often affected by variations in the scene itself and as well as the setting parameters (illumination, focus, motion, etc.) of the used camera.



Figure 2: Scene text

1.2 Tools and Algorithms

Will be used algorithms for each phase of extracting text from a video using java libraries and classes. Here first we frame the input video into stream of images using the Java Media Framework (JMF). Then we apply pre processing algorithms to convert the image to gray scale and remove the disturbances like superimposed lines over the text, discontinuity removal, and dot removal. Then we continue with the algorithms for segmentation for which we use the fuzzy c-means technique.

1.2.1 Canny detector

Canny detector which consider region- based detection. canny Steps

- a) Smooth image with Gaussian.
- b) Compute gradient and angle images.
- c) Apply non-maxima suppression to M.
- d) Use double thresholding and connectivity analysis to detect and link edges [5]

1.2.2 fuzzy c-means

Fuzzy K-Means (also called Fuzzy C-Means) is an extension of K-Means, the popular simple clustering technique. While K-Means discovers hard clusters (a point belong to only one cluster), Fuzzy K-Means is a more statistically formalized

method and discovers soft clusters where a particular point can belong to more than one cluster with certain probability. following equation show fuzzy c-means algorithm:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty \quad 1$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i th of d-dimensional measured data, c_j is the d-dimension center of the cluster. and $\|*\|$ is any norm expressing the similarity between any measured data and the center, I is text cluster and j is background cluster[6]

In this paper, we propose a new semantic video-text-detection approach that is capable of detecting text in a complex background, scene text. Section 2 Overviews previous work.3 AFS arch Section 4 Semantic enhancement of TIE Section 5 draws our conclusions and future work.

2. Previous Work

Previous work can be divided into two categories as following:

- a. Multimodal such as k-space.
- b. Text information extraction such as Region-based and Texture-based.

The K-Space project deals among others with the integration of knowledge structures, as encoded in high-level representation languages, and low-level descriptors for audio-video content, taking also into account knowledge that can be extracted from sources that are complementary to the audio/video stream, mainly speech transcripts and text surrounding images or textual metadata describing a video or images, or even text included in the images. These complementary resources typically fall into two groups: primary resources, which are directly attached to multimedia, and secondary resources, which are more loosely coupled with the audio-video material.[7]

In the literature, text detection and localization are regarded as one problem. This section reviews the main methods proposed to deal with text detection/localization in images and videos. Jung, have classified text detection/localization methods according to the employed features into two main groups: region-based and texture-based [2]. The first class employs connected component analysis. It is

based on the analysis of the geometrical arrangement of edges or homogeneous color and grayscale components that belong to characters [8]. While, these methods are simple to implement, they are not very robust for text localization in images with complex background. The performance of these methods depends on several values such as Size, alignment and Color.

Chen et al. [8] used the Canny operator to detect edges in an image. Only one edge point in a small window is used in the estimation of scale and orientation to reduce the computational complexity. The edges of the text are then enhanced using this scale information. Morphological dilation is performed to connect the edges into clusters. Some heuristic knowledge, such as the horizontal-vertical aspect ratio and height, is used to filter out non-text clusters. Two groups of Gabor-type asymmetric filters are used for generating input features for scale estimation: an edge-form filter and a stripe-form filter and a neural network are used to estimate the scale of the edge pixels based on these filters' outputs. The edge information is then enhanced at an appropriate scale. As such, this results in the elimination or blurring of structures that do not have the specified scales. The text localization is applied to the enhanced image. The authors used a commercial OCR package after size normalization of individual characters into 128 pixels using bilinear interpolation.

The second class of methods, consider text as regions with distinct textural properties [9]. Whereas these methods are more accurate, when the text is embedded in a complex background, they have difficulties in detecting text of different sizes and they are computationally expensive.

Sin et al. [10] use frequency features such as the number of edge pixels in horizontal and vertical directions and Fourier spectrum to detect text regions in real scene images. Based on the assumption that many text regions are on a rectangular background, rectangle search is then performed by detecting edges, followed by the Hough transform. However, it is not clear how these three stages are merged to generate the final result.

AFS will use Canny detector which consider region- based detection (second class).

3. AFS Architecture

At this stage will be explained the processes and algorithms for each stage of the AFS framework which can be divided into two parts the first is extraction of text from video, Part II extract semantic by using wordNet library then save result in meta data file.

Figure 3, illustrates the stages of the framework proposed for text extraction from a given video clip, each of which has its role in dissolving the problem under study.

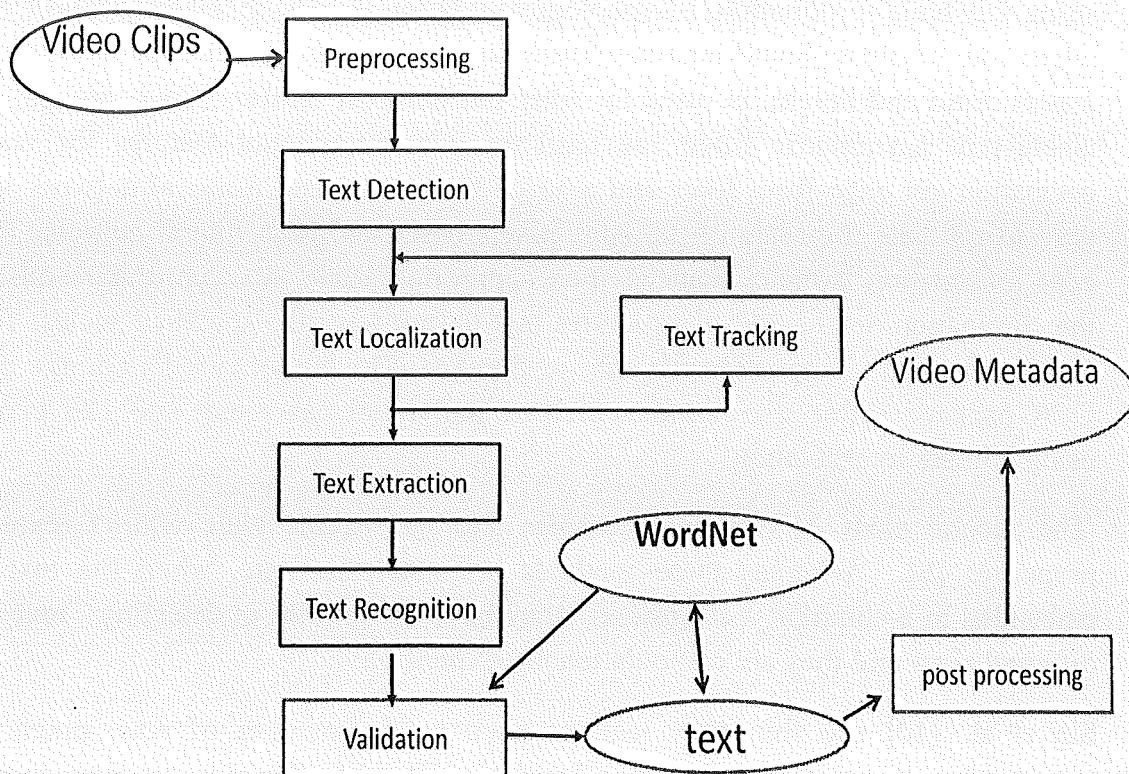


Figure 3: AFS Architecture

- Preprocessing Stage: it is responsible for conversion and segmentation of the given clip.
 - a) Convert video files into the mpeg standard type,
 - b) Segment the video into images.

- Text Detection stage: it determines the existence or nonexistence of text in the given image. However, number of frames containing text in, most video clips, are much smaller than those without text.
- Text Localization: it is responsible for grouping text regions into text instances and generating a set of tight bounding boxes around all text instances. It employs region based methods for text localization. Region based methods relies on the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background.

The text region localization task is achieved in two steps, the extraction of a candidate text block, followed by the extraction of individual text line.

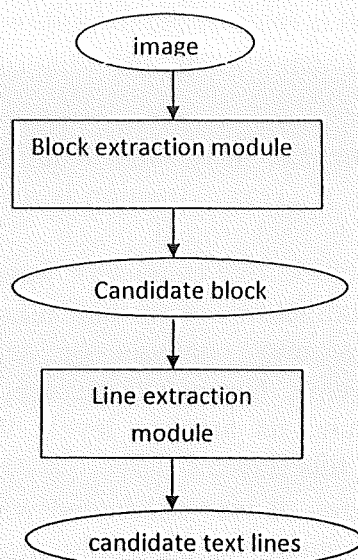


Figure 4: Text localization sub-tasks

In Table 1, Chethan et.al, summarized their findings regarding the performance of four of the edge based detection algorithm [11]. Accordingly, the Canny edge detection² algorithm was adopted for application in text localization.

² known as the optimal edge detector

Table 1: Comparison of different Edge based Methods [11]

No	Type of images	Recognition Rate (%)Canny	Recognition Rate (%)SOBEL	Recognition Rate (%)Log Transform	Recognition Rate (%)Prewitt
1	Clean	99.2	79.6	95	76
2	Blur	75	68	72	62
3	Orientation	79	79	65	70
4	Illumination	80	64	76	65

- **Text Tracking Stage:** it is responsible for following a text event as it moves or changes over time and determining the temporal and spatial locations and extents of text events.
- **Text Extraction:** After the text was localized, the text extraction step deals with the text box's. Text extraction is responsible for the separation of the text pixels from the background pixels. The output of this step is a binary image, where, black text characters appear on a white background. This stage included extraction of actual text regions by dividing pixels with similar properties into contours or segments and discarding the redundant portions of frame. Through this stage, text is marked as one binary level and the background as another level. Will be used fuzzy c-means algorithm to completion this stage.
- **Text Recognition Stage:** it was conducted with commercial OCR engines. The traditional OCR technique handles, only, those binary text image that have simple backgrounds. while the images colored images which have complex background will contain a lot of noise and disturbances. After choosing the most likely letters in a given word, the word can be compared with a dictionary lookup (WordNet), to verify if the letter combination is likely.
- **Generating Metadata:** will be using a newer version of Adobe Creative Suite it includes an application called Bridge which is "the control center for Adobe Creative Suite." This application enables file organization, etc. It also allows for the addition of metadata on certain files. In particular it is able to do so with AVI and MOV files, but we were not able to add metadata to an MP4. *Bridge is*

an application of *Adobe Creative Suite*. It was adopted to generate and store the metadata in an accessible format.

Adobe Bridge creates the data in Extensible Metadata Platform XMP, which is based on XML and was created by Adobe Systems. This has essentially become the de facto standard for most embedded metadata. It allows the metadata to be stored in the file without making the file unreadable by non-XMP support applications.[12]

4. Enhancement of *TIE*

Will be in the following paragraphs explained the new components (wordnet library and metadata file) added to *TIE* framework in AFS as described in figure 4.

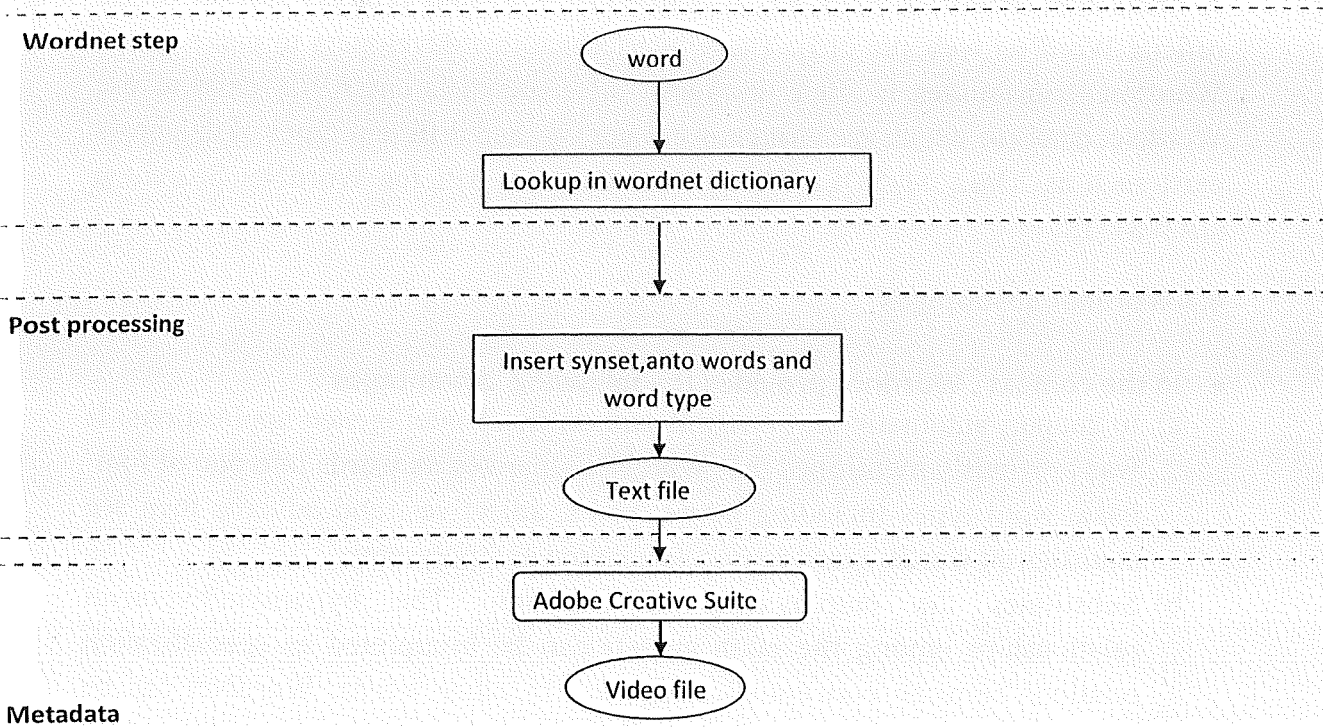


Figure 5 :New AFS processes

4.1 Employed WordNet

WordNet is employed to organize the words and their relationships into a hierarchical structure similar to that of taxonomies, which may be, found in the natural sciences.

Words which are closely related to each other may be found in the same branch of

the tree hierarchy. Each word belongs to a set of synonyms, known as a synset. Formally, a synset is a set of one or more synonymous words that may be substituted for each other in context without changing the overall meaning of the sentence in which they were included. Words which have multiple meanings or “word senses” are contained in more than one synset.

WordNet provides a polysemy count for each word to track the number of synsets which contain the word. It includes a frequency score which uses a sample text to determine how often a word appears in a specific sense. The database is able to track semantic relationships connecting synsets that are specific to each word type, as well as lexical relationships between individual words [13], [14], [15].

4.2 Developing with WordNet

A number of WordNet APIs are available in the related projects section of the official WordNet web site [15]. These interfaces allow developers to make use of WordNet's capabilities in their programs. there are different API's exist to use WordNet database which are written in Java or C#, and different types of databases exist such as XML, MySQL. We will use the java API for WordNet Searching (JAWS) [13].

4.3 Wordnet advantages

- **In the context , Wordnet provides the following advantages :**
 - a) to support automatic text analysis and artificial intelligence applications: It was previously mentioned that WordNet separates its words into one of four basic word types including nouns, verbs, adjectives, and adverbs then see statistics on words, word senses, synonymy and polysemy.
 - b) Substitutes the employment of ontology: The hypernym/hyponym relationships among the noun synsets can be interpreted as specialization relations between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology in the computer science sense.
 - c) Validates OCR results.
 - d) Can be mapped between the wordnet and the other ontology such as upper ontology sumo.

4.4 Post-Processing

This stage is responsible for exploring the words and the lines of given video frame and generate information for generate the metadata as follows:

- Generate a text file 1 that contains the semantics and ontologies information about each words resulting from TIE.
- A group of words that constitute a line in the video image is processed into text file 2.this file contains the video frame number and text region.
- Merge two file on one file.

5- Results

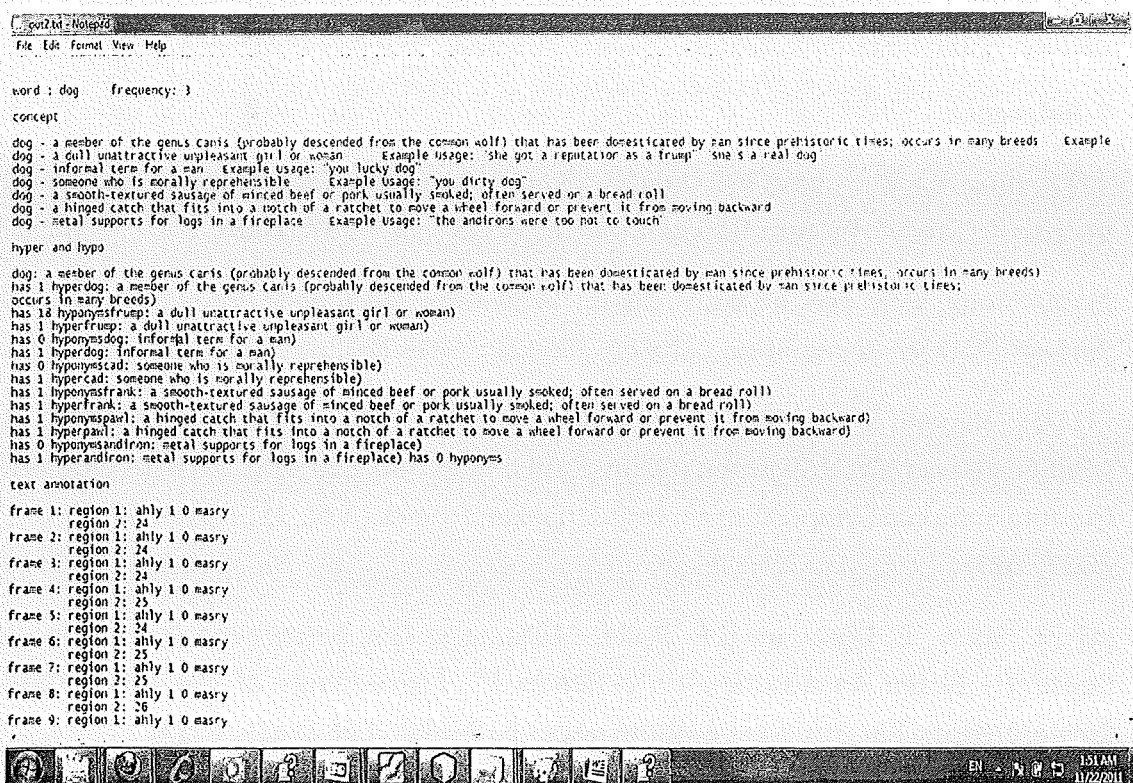


Figure 5: Final text file

Fig five shows AFS result (final text file) which can divide into two parts. First part, appear at the top contains the semantics and ontologies information about each words resulting from TIE. Second part, which contains the video annotation for each frame in video file.

6-Conclusion and Future work

In this Paper, we have proposed an enhanced approach to AFS for text detection in videos supported by WordNet library. Using wordnet aims to enhance extracting semantic from video text by Using wordnet library as secondary complementary resource (ontology).

Use the WortNet instead of ontology provides AFS to work with more domains (Sports, news, lectures videos). AFS metadata file storage at video file, this metadata can use by multiple applications such as automatic search engines or automation summarizing application.

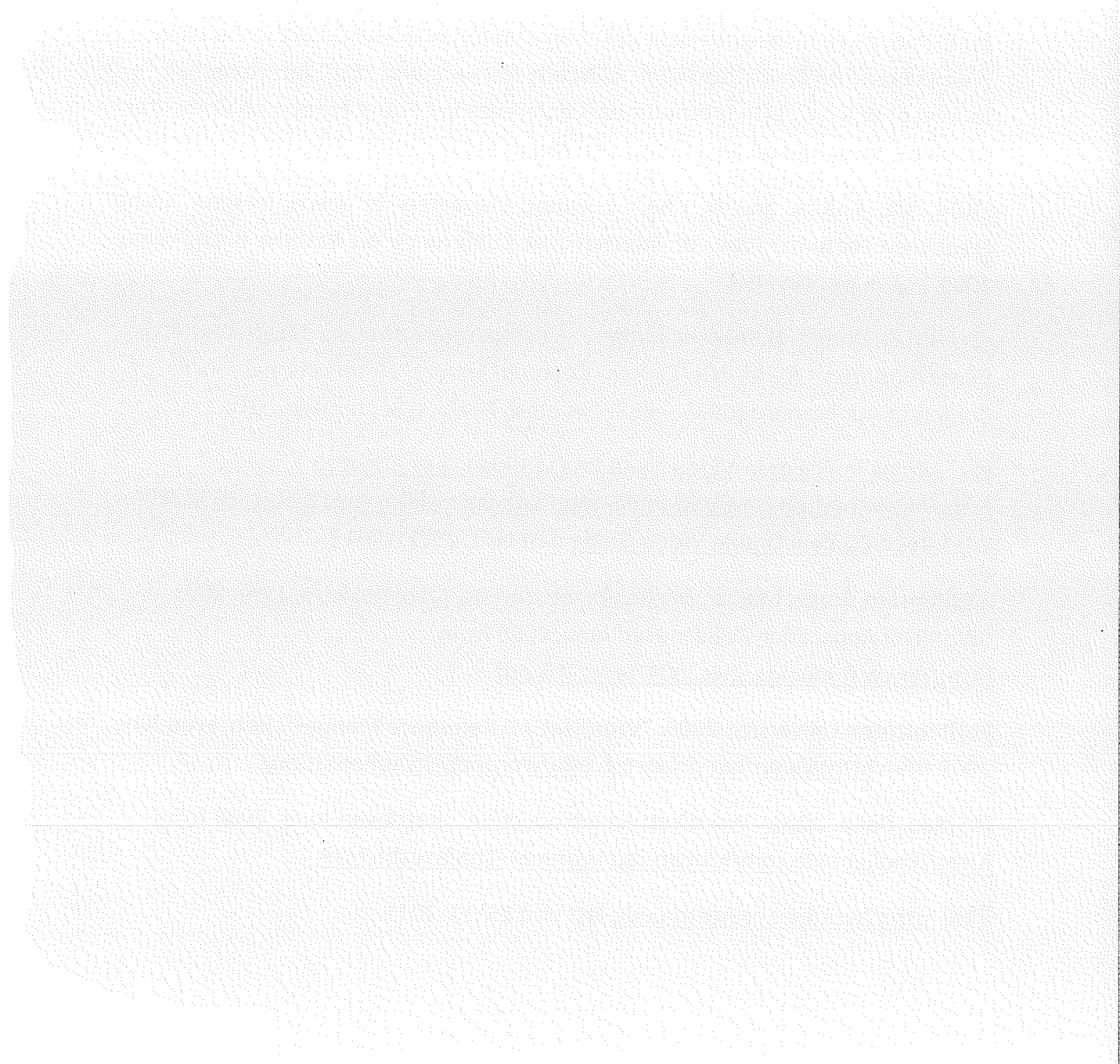
Future work is to integrate the AFS framework into the multimodal framework include of the rest of the complementary resource such as audio and face detection.

7-Reference

- [1] K. Jung, K.I. Kim , an d A.K. Jain . "*Text In form action Extract ion in Images and Videos: A Survey*". Pattern Recognition Letters ,37:977–997, 2004.
- [2] Y.Y. Tang, S .W. Lee, an d C.Y. Suen . "*Automatic Document Processing: A Survey*". Pattern Recognition Letters , 29(12) :1931–1952, 1996.
- [3] D. S . Kimand S . I. Chien . "*Automatic Car License Plate Extraction using Modified Generalized Symmetry Transform and Image Warping*". Proceedings of the International Symposium on Industrial Electronics (Vol. 3) , p ages 2022–2027, 2001.
- [4] J.C. Shim , C. Dorai, an d R. Bolle. "*Automatic Text Extract ion from Video for Content - Based An notation an d Retrieval*". Proceedings of the International Conference on Pat tern Recognition ,pages 618–620. IEEE Press, 1998.
- [5] <http://www.altera.com/literature/cp/gsp/edge-detection.pdf>. last visit: 25\11\2011.
- [6] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html. Last visit: 25\11\2011.
- [7] <http://www.altera.com/literature/cp/gsp/edge-detection.pdf>. last visit: 25\11\2011.

- [8] D. Chen , K. Shearer , and H. Bourlard . "*Text Enhancement with Asymmetric Filter for Video OCR*". Proceedings of the International Conference on Image Analysis and Processing, Pages192–197. IEEE Computer Society, 2001.
- [9] Y. Hao, Z . Yi, H. Zengguang, and T. Min . "*Automatic Text Detection in Video Frames based on Bootstrap Artificial Neural Network and CED*". International Journal of W S C G, 11th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 11(1) , 2003.
- [10]B. Sin, S. Kim, and B. Cho, Locating Characters in Scene Images using Frequency Features, Proc. of International Conference on Pattern Recognition, 2002, Vol. 3, pp. 489-492.
- [11] H.K.Chethan,G.Hemantha Kumar." *A Comparative Analysis of Different Edge Based Algorithms for Mobile/Camera Captured Images*". International Journal of Computer Applications (0975 – 8887) Volume 7– No.3, September 2010.
- [12] Adding In-File or In-Video Metadata to Video Files - How to
<http://www.reelseo.com/add-infile-metadata-video-files/#ixzz1ZzfEcFFR> ©2008-2011 ReelSEO.com Online Video Guide .last visit: 25\11\2011.
- [13]Morato, Jorge; Marzal, Miguel Ángel; Lloréns, Juan; Moreira, José. 2003. "WordNet Applications". Retrieved June, 2009 from <http://www.fi.muni.cz/gwc2004/proc/105.pdf>.
- [14]Princeton University. 2009. "WordNet 3.0 Reference Manual". Retrieved June, 2009 from <http://wordnet.princeton.edu/wordnet/documentation/>.
- [15]Rao, Delip. 2008. "WordNet: An Introduction". Retrieved June, 2009 from <http://knol.google.com/k/delip-rao/wordnet/1tipfaxusbh7/3#>.
- [16] <http://wordnet.princeton.edu>.last visit 25\11\2011.

THE UNIVERSITY OF CHICAGO
LIBRARY



Cloud Community Framework for Web Services Using Ontology

Tamer M. Al-Mashat¹

Fatma A. El-Licy¹

Akram I. Salah²

Abstract

Cloud computing describes a broad movement toward the use of wide area networks (WANs), such as the Internet. Cloud computing facilitates and enables interaction between information technology (IT) service providers and consumers. Service providers are expanding their service to include software applications. In this paper the quest is to develop cloud community framework architecture for web service to cater services for end users. The proposed cloud community framework combines the web services and its semantics with the cloud paradigm to reform the internet into groups of community hierarchies. An upper community concept is introduced with a general ontology for controlling and categorizing communities. Service quality measures are introduced for nonfunctional web service. This quality measures assist cloud community and end users in discovering the preferences services .The framework is a conceptual model for the web services to make it clearly understood and efficiently outsourced.

Keywords: Cloud community, Software as a Service (SaaS), web service, ontology, service discovery, service composition

1. Introduction

It seems probable that, eventually, most software capabilities will be delivered and consumed as services [1]. Software as a service (SaaS) is one of the methodologies of Cloud Computing. SaaS is a way of delivering applications over the Internet as a service [2]. SaaS has a lot of benefits for both the consumer (end users) as well as the providers. SaaS enables customers to access and use software over the internet without the complexity of managing and installing it. Not only service providers gain economic benefits, but also they get feedback from customers which is reflected on the improvement of the services provided. Figure 1 consolidates some of these benefits realized in a set of interacted points.

¹ Institute of Statistical Studies and Research Computer Science Department Cairo University

² Faculty of Computer science and Information Computer Science Department Cairo University

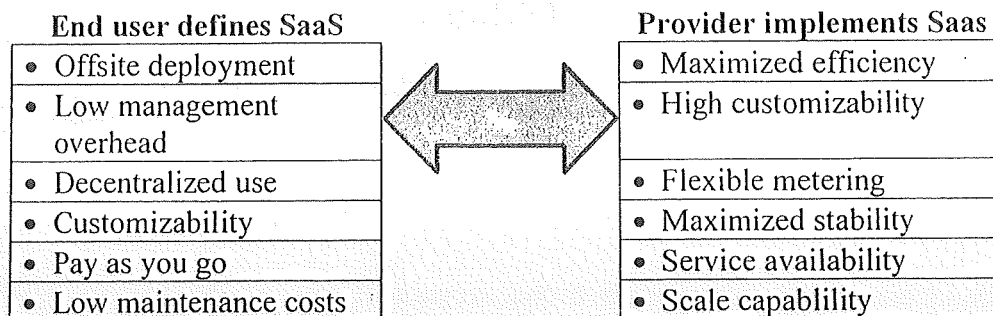


Figure 1: Benefits of SaaS

Realizing benefits of SaaS necessitates shifts in thinking approach on the part of both the provider and the customer.

The growing trend in software architecture is to build platform-independent software components, called Web services that are available in the distributed environment of the Internet. More complex applications are to be assembled from a set of appropriate Web services rather than be written manually [3]. Seamless composition of Web services has enormous potential in streamlining business-to-business or enterprise application integration. Business process can encapsulate their functionality as web services and create virtual process.

Business domains have requirement for automatically discovering and composing web services according to specified business logic. Therefore, they require an efficient method of selecting web service according to their component requirement. Web service composition addresses the situation when a client request cannot be satisfied by any available web service, but by suitably combining “parts of” available web services, the client request can be achieved.

The main challenge lies in the massive number of web services and the difficulties to determine validity of service provider. There is a need for technique to organize and categorize Web services reference to a specific domain of interest

The Web services are, usually, developed by different organizations that apply different technology and conceptual models for presenting services characteristics. This requires utilization of relevant semantic information of Web services for facilitating of web services operations and integrations. It is hard to unify concepts between service providers even if they were working at the same domain. Therefore, a method is required to unify the concepts, in a specific domain, between service providers. Concept unification will maximize the role of semantics information.

This paper introduces the concept of cloud community, which is a set of service providers working in a common domain of interest and introduce services over the internet. The introduction of cloud community provided for representing shared concepts of a specific domain in an ontology. This leads to facilitate an automation of web service discovery and composition to provide services for end user. The cloud community, also, services as an evaluation entity for quality of service to provide trusted services. Simply, cloud community adds a level of abstract to the cloud setting, a semantic level to understand the context. This paper defines a framework for service provision based on cloud community.

The rest of this paper is organized as follows: Section 2, reviews web service architecture, in which web service is a main building block of service oriented architecture and semantic web services for enhancing its operations and integrations. Section 3, discusses the proposed cloud community framework. Section 4, presents a method for facilitating interoperability between different cloud communities based on exchanging semantics information. Section 5, introduces criteria for evaluating quality of web services.

2. Web Service Architecture

Service oriented architecture (SOA) is a set of components which can be invoked, and whose interface descriptions can be published and discovered [4]. Web Services currently provide the main enabling technique for SOA. In fact web services are not a mandatory component of a SOA, but web services characterized by various industry support, ubiquity, and easy understood concept.

Web services are self-contained, self-describing and modular applications. Because a Web services adopt open standard interfaces and protocols, they are increasingly, used to integrate and build business applications over the Internet [5]. Web services in general use of client server interaction using XML messages that follow a common Simple Object Access Protocol (SOAP)³ standard protocol. Web Services are platform independent and can work on any programming languages.

2.1 Web Service Based Architecture

Figure 2, illustrates the main components of web service architecture and their interrelationships. The main components of typical web service based architecture are Service requestor, Service provider and Service broker. The service requestor is the client and service provider is the server. The service broker acts as a repository of services. The service provider publishes its services on the service registry. The service requestor can use the service published by the service provider by discovering the service via the service

³<http://www.w3.org/TR/soap12-part1/>

registry. If it is available then the registry replies back with the service information. Then the service requestor binds a connection with service provider and executes the service [6].

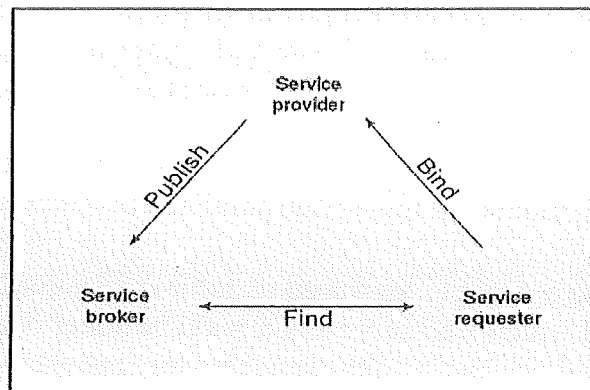


Figure 2: Web Service Architecture

The specification of a web service is expressed in web service description language (WSDL) [7]. WSDL provides a function-centric description of Web services covering inputs, outputs, and exception handling. The Semantic Web provides a process level description of the service which, in addition to functional information, models the preconditions and post conditions of the process so that the evolution of the domain can be logically inferred. It relies on ontologies to formalize domain concepts which are shared among services.

3. Cloud Community Framework Based On Ontology For Web Services

The large scale, dynamics, and heterogeneity of Web services may conceal any attempt for understanding their semantics. These lead to difficulties of web services discovery and composition automation [8]. These difficulties call for techniques to organize Web services in a way they can be efficiently understood and outsourced.

3.1 Cloud Community

Figure 3, illustrates web service architecture based on a cloud. The internet is divided into groups of communities' reference for a common of interest [9]. A community is a cluster of web services that reference for a domain of interest. A cloud community caters a description of web services and ranking them based on the metric measures and feedback of services consumers.

Community providers are generally groups of non-profit organizations, and businesses that share a common domain of interest and introduce services on the internet. Communities are published in an upper community registry so that they can be discovered by service providers, identify the community of interest and

register their services with it. Service requesters identify the community of interest and demand the services.

The main objective of a community is to provide services to consumers. These services may be mapped into one or more web services to implement them; in the last case we need a tool to automate web services composition.

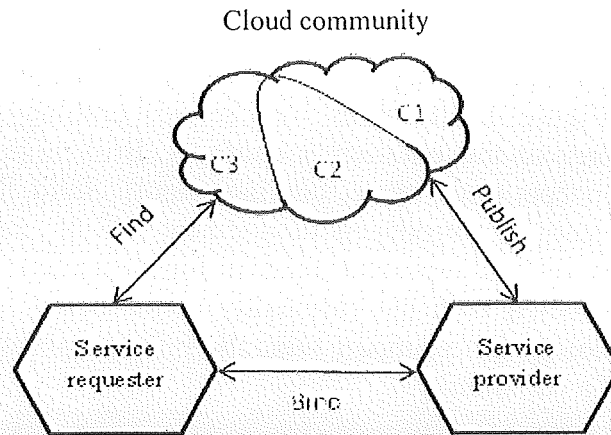


Figure 3: Web Service Architecture based on cloud

3.2 Services provided by the Cloud Community

A cloud community is mediation between, service providers and service consumers, caters a set of services as follow:

- A cloud community caters web services interfaces for services providers to implement them.
- Unify Web Services Standards and Specifications.
- Unify concepts for a specific domain.
- Unify concepts for a business process.
- Unify annotation method for web services.
- Ranking web services according to feedback of end user and quality measures.
- Unify business process workflow according to a cohesive set of best practices and standard; automate business process according to these best practices.
- Provides a collection of agents as services to co-operate together for automating execution of services.

3.3 Cloud Community Ontology For Web Service

Semantic Web technology, namely ontology⁴ languages, provides scalable methods and tools for the machine-accessible representation and manipulation of knowledge. Semantic Web Services (SWS)⁵ make use of Semantic Web technology to support the automated discovery, substitution, composition, and execution of web services [10].

A key for creating a framework for dynamic business processes is semantically representing the functional and non-functional capabilities of web services. Figure 4, illustrates web service ontology using a directed graph. Nodes represent the ontology's concepts. Broken line ovals (nodes) refer to WSDL concepts (e.g., name, binding, input). Solid line ovals (nodes) refer to extended features introduced to augment WSDL descriptions with semantic capabilities. Edges represent relationships between the ontology's concepts. They are labeled with the cardinality of the corresponding relationship. Web service ontology is to support semantic interoperability between a web services on level of a community.

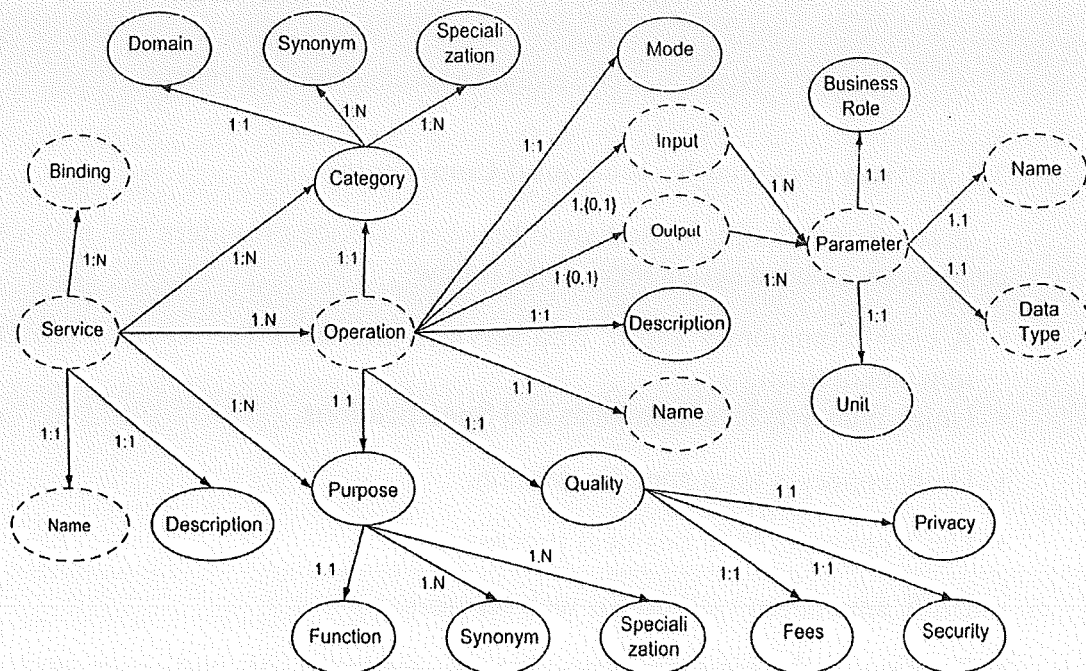


Figure 4: Web service ontology[11]

3.4 Cloud Community Architecture Framework

When a service consumer demands a service from a community, the request handler receives and transmits it to the security and policy layer to be checked. Security and policy layer represent a gate for the community core. An Agent container is a collection of agents who are co-operating together to fulfill a

⁴an ontology is a formal, explicit specification of a shared conceptualization

⁵<http://www.w3.org/Submission/OWL-S/>

given requested services. These are, normally, provided by the community. Managed communication infrastructure is a layer for controlling and managing interaction between parties of the community. Business process template repository is a container for business process workflow. A workflow was, usually, selected according to the standard and best practice. Semantic business process is standard definition of concepts, utilized to express business process models (a metamodel). Metamodels define concepts, relationships, and semantics.

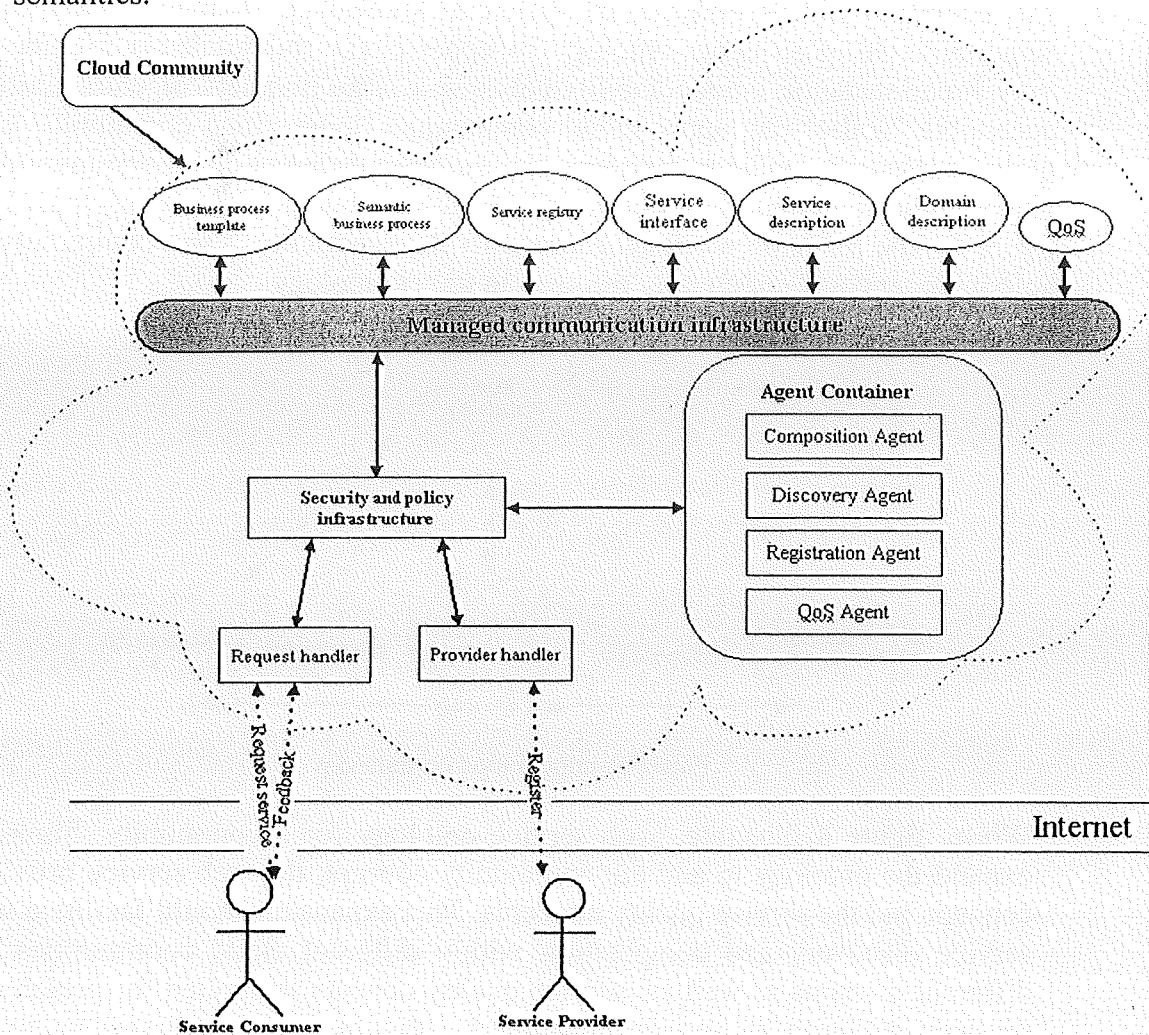


Figure 5: Cloud Community Architecture Framework

Service interface repository is a collection for interfaces of web services. The interfaces are, normally, provided by the community provider and implemented by a service provider. Service repository represents a collection of web services, whereas, service description is a repository for its semantic. Domain description repository is a collection of ontologies for both, community domain and its subdomains. QoS repository is a reference for web service quality. Each service

in the web service repository is mapped into quality repository to represent a record of some evaluated criteria of quality measured.

4. Interoperability Among Cloud Communities

Upper cloud community facilitates interoperability between different communities as shown in figure 6. It provides global ontology to support cloud communities of web services. Upper ontology is an ontology which describes the concepts with synonyms in different knowledge domains. The most important role of an upper ontology of upper cloud community is to support very broad semantic interoperability between a large numbers of communities accessible "under" this upper cloud community. Upper cloud community is a global community orchestrates semantics between different cloud communities. Upper cloud community is a community of other cloud community as shown in figure 6. The semantic of cloud community 1, community 2 and community 3 hold in upper cloud community through an upper ontology.

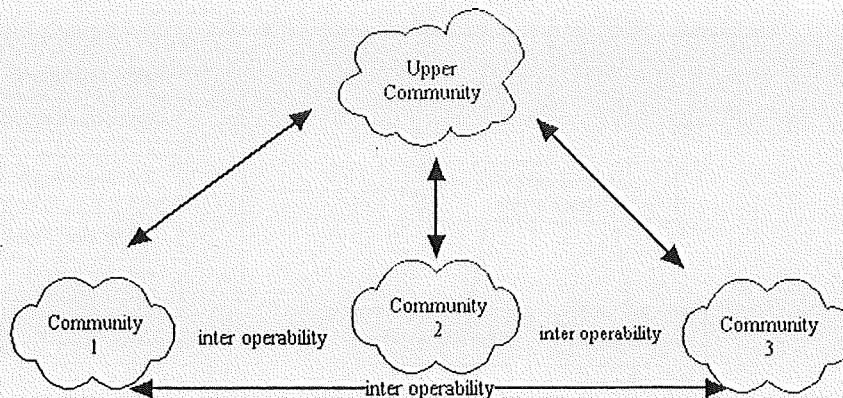


Figure 6: upper cloud community

While, figure 4 illustrates web service ontology for a community. Figure 7 expand the ontology to add the community concept, the web service provider concept and the service quality concepts. This expansion should facilitate web service discovery and composition within different cloud communities for different domains as well as, service quality concept is applied for preference services selection by service consumers. Figure 7, demonstrates an upper ontology for all cloud communities of web services. Multiple ontologies of cloud communities will co-exist to build a global ontology of upper cloud community. This upper ontology could be a template for describing communities and web services beneath the upper cloud community which facilitates interoperability between cloud communities. In addition, upper ontology will be as index for web services to accelerate service discovery.

These measures will be calculated statistically according to feed back of end users and evaluation of the community provider. Table 1, presents a number of criteria for quality of services (QoS), in the context of individual web service. These criteria help end user to, optimally, select component of services according to his/her preference.

As mentioned previously, a cloud community caters outsourcing trusted services for end user. Registering a web service with a cloud community refer to the process of importing web services from service providers. The invocation of an imported web services is translated into the execution of an actual services demanded by end users. In addition, a cloud community provides standards and domain-specific ontology which facilitate web services discovery and composition.

6. Conclusion

This paper proposed a cloud community framework architecture that provides a meaningful organization and description of web services. The concept of community is introduced to cluster web services based on their domain of interest. Upper community ontology was developed to serve as a template for describing communities and web services. The ontology based cloud community framework lays the foundation for the automatic selection and composition of semantic web services.

References

- [1] Ruehl S. and andelfinger U., "Applying software product lines to create customizable software-as-a-service applications" *Proceedings of the 15th International Software Product Line Conference*, Vol. 2, Aug., 2011.
- [2] Armbrust M., Fox A., Griffith R., Joseph, Katz A., Konwinski A., Lee G., Patterson D., Rabkin A., Stoica I., and Zaharia M., "Above the Clouds: A Berkeley view of Cloud Computing," *University of California*, Berkeley, 2009.
- [3] Jiangang Ma., Yanchun Z., Minglu L., "OMWSC- An Ontology-Based Model for Web Services Composition," *Fifth International Conference on Quality Software*, 2005.
- [4] Sprott D., Wilkes L. "Understanding Service-Oriented Architecture"

<http://msdn.microsoft.com/en-us/library/aa480021.aspx>.

Last visit 30/11/2011

- [5] Curbera F., Khalaf R., Mukhi N., Tai S. and Weerawarana S.,“ The next step in Web Services”,*Communications of the ACM*, 46(10), pp. 29-34,2003.
- [6] Gottschalk K. and IBM Team. 2000. Web services architecture overview: The next stage of evolution for e-business. Article, IBM,
<http://www-106.ibm.com/developerworks/web/library/w-ovr/>.Last visit 30/11/2011
- [7] Christensen E., Curbera F., Meredith G. and Weerawarana, S., ” Web-Services Description Language (WSDL),” 2001. <http://www.w3.org/TR/wsdl>. Last visit 30/11/2011
- [8] Milanovic N. and Malek M., “Current solutions for Web service composition”, *IEEE Internet Computing*, 8(6), pp.51 – 59, 2004.
- [9] Marinos A. and Briscoe G. “community cloud computing”, *First International Conference on Cloud Computing*, 2009.
- [10] OWL-S 1.0 Release,
<http://www.daml.org/services/owl-s/1.0/>. Last visit 30/11/2011
- [11] Medjahed B., Bouguettaya A. and Elmagarmid A., “Composing Web services on the Semantic Web” *The VLDB Journal*, pp: 333–351, 2003.

إطار عمل الحوسبة السحابية لمجتمع خدمات الويب باستخدام خريطة المفاهيم

الحوسبة السحابية تظهر إتجاه واسع نحو إستخدام شبكه الانترنت لتسهيل التفاعل بين مقدمي خدمات تكنولوجيا المعلومات ومستهلكين هذه الخدمات . مقدمي خدمات الانترنت يقومون بتوسيع نطاق خدماتهم لتشمل تطبيقات البرمجيات. تسعى ورقة العمل هذه الي تطوير بيئة بنية سحابية لاطار المجتمع الخدمي للويب لتلبية الخدمات المطلوبة من المستخدمين. إطار بيئة عمل الحوسبة السحابية للمجتمع المقترحة تجمع بين خدمات الويب ومعانيها الدلالية مع نمذجه الحوسبة السحابية لتقسيم شبكة الانترنت إلى مجموعات. في هذه الورقة يتم عرض مفهوم المجتمع العلوي مع خريطة المفاهيم العامة التي تساعد في تصنيف وتعاون المجتمعات فيما بينها . بالاضافة الي وضع عدد من المعايير لاختيار هذه الخدمات مما يساعد المستخدمين في اختيار هذه الخدمات وفقا لهذه المعايير.

Rhetorical Representation for Arabic Text

Ahmed Ibrahim ¹

Tarek Elghazaly ²

Abstract

Rhetorical Structure Theory (RST) is centered in different areas of science, like discourse analysis, theoretical linguistics, psycholinguistics, and computational linguistics (information retrieval, text generation, text summarization and machine translation). RST is applied in English, and in other languages, such as French and Japanese but there is no clear effort has been spent to apply RST on Arabic language which hundreds of millions of people are speaking. This descriptive framework can analyze discourse structure and provides a rhetorical representation like a binary tree for Arabic text that improve the ability of extracting nucleus node that represent more important parts of text.

Keywords

Rhetorical structure theory, rhetorical structure tree, Arabic rhetorical structure, RST.

1. Introduction

Rhetorical Structure Theory³ (RST) is descriptive theory of a major aspect of the organization of natural text, it's a linguistically useful method for describing natural texts, characterizing their structure primarily in terms of *relations* that identify particular relationships that can hold between two non-overlapping *text spans*. Depending on relation type, the text span can be *nucleus* or *satellites*, nucleus are considered the master seed and importance parts of a text, whereas *satellites*

¹ Ahmed Ibrahim master student, Institute of Statistical Studies and Research, Cairo University.

² Dr.Tarek Elghazaly, Institute of Statistical Studies and Research, Cairo University.

³ The name Rhetorical Structure Theory has been criticized in the literature by three groups. One complains that Rhetorical is never related to classical rhetoric. A second suggests that a non-standard meaning of Structure is being used. The third group points out that RST is thoroughly descriptive and does not put forth a causal Theory. All of these are commendable, valid arguments. The name of RST was chosen as a phrasal unit, distinctively best out of a collection of suggestions [11].

contribute to the nucleus and are secondary. For instance, in an Elaboration relation, the nucleus is considered to be the basic information, and the satellite contains additional information about the nucleus. The nucleus is more essential to the writer's purpose than the satellites. The satellites are often incomprehensible without the nucleus, whereas a text where the satellites have been deleted can be understood to a certain extent [1] [2].

2. Prior work

Rhetorical Structure Theory Originated initially in 1980s at the Information Sciences Institute (ISI) of the University of Southern California (USC) by William C. Mann with major focus on natural language generation and on developing the discourse linguistics necessary computer generation of natural language text. It was intended for a particular kind of use to computational text generation. He started work with Sandra A. Thompson as Consultant on discourse (ISI) [3] [4].

Marcu [5] [6] [7] gives more detailed description of the rhetorical parsing strategy, the rhetorical parsing, text summarization, and text generation of natural language.

Relation definitions are based on functional and semantic criteria like set of cue phrases. The original set defined in the Mann and Thompson paper includes 24 relations², called "Classical RST".

Last set in Arabic by Alsanie [4] defines 11 relations with cue phrases as table1, used for text the summarization and identifying nucleus parts that considers more important points for text summarization.

3. Arabic rhetorical relations

Arabic grammarians were interested in studying analytical and not synthesis for Arabic sentences therefore, the results were lacking in a comprehensive synthetic theory of Arabic sentences in general. Their study was far from the comprehensive view because the idea of relations was not part of their approach and therefore did not list all the

² Twenty-three relations proper, and one schema, Joint.

Table1: Arabic Relation with Cue Phrases [4].

<i>Arabic Rhetorical Relation</i>	<i>Name in English</i>	<i>Cue-Phrase</i>
شرط	<i>Condition</i>	لو- ان
تفسير	<i>Interpretation</i>	بمعنى - حيث
تعليق	<i>Justification</i>	بسبب - لان
استدراك	<i>Recalling</i>	لئن
توكيد	<i>Confirmation</i>	أذن - بل
نتيجة	<i>Result</i>	لذا - نتيجة لذلك - نستنتج
تمثيل	<i>Example</i>	مثالا - على سبيل المثال
قاعدة	<i>Base</i>	بما ان - بناءا على
تفصيل	<i>Explanation</i>	الا ان - غير ان - لاسيما
عطف	<i>Joint</i>	و
ترتيب	<i>Sequence</i>	ثم - من ثم

Arabic rhetorical relations in general [8]. This led to the delay of implementing Arabic RST until 2005 [4], as identified in the relations shown in Table 1.

This paper hypothesizes Arabic Rhetorical Relations that collected from RST applications of Arabic [4] and Arabic literature references [8], to get the full set of rhetorical relations as clearly defined, which can be used within this framework.

4. Verification of the Arabic Rhetorical Relations Hypothesis

The verification of an Arabic rhetorical relation hypothesis is applied on Arabic corpus, to see if hypothesis relations are explicitly signaled or not. This requires developing an Arabic corpus by using an HTML parser module for extracting the completed articles from the different Arabic news portals. Then identifying the article structure, i.e. article title, summary if found and article paragraphs by the use of HTML tags.

The proposed framework matches the first cue on every paragraph with the cues of the rhetorical relation hypothesis in order to define the relationship between paragraphs and identify nucleus and satellite paragraphs.

Classify rhetorical relations based on the frequency occurrence on corpus. Figure 1 illustrates the joint (العطف) relation that takes the first order by 76% because the joint relation contains conjunctions like ("and") ("واو العطف") that have a higher frequency occurrence of up to 1015 times, as in Figure 2. All these cues can be used to define the satellites paragraphs.

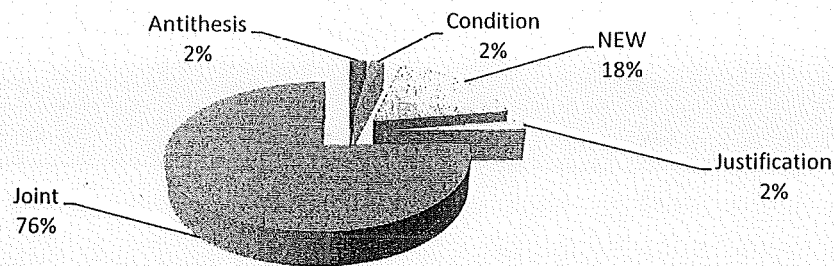


Figure 1: A Statistical relations frequently occurrence.

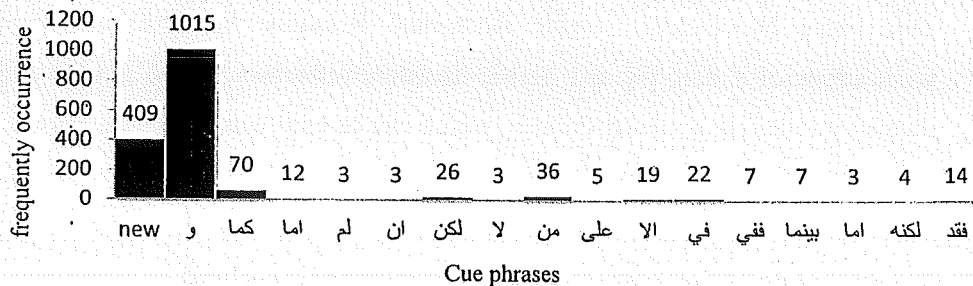


Figure 2: Statistical cue phrase frequent occurrence for satellite paragraphs.

The second order frequency occurrence relation is 18% for a new relation. This new relation includes cues that are not identified in the matching phase. All these cues are shown in Figure 3 and can be used to define the nucleus paragraphs.

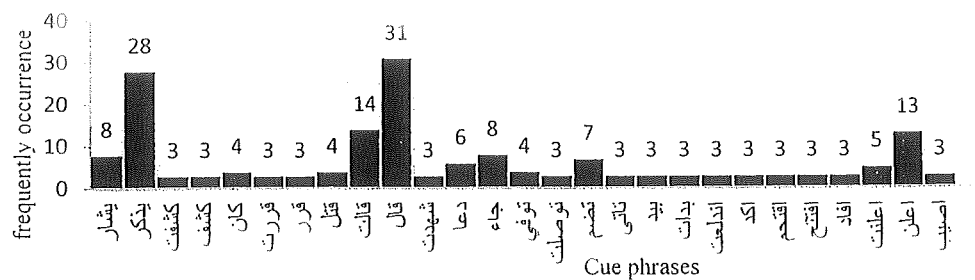


Figure 3: Statistical cues frequent occurrence for nucleus paragraphs.

The statistical reports are manually reviewed to discover cues that are included in improper relations. To discover these cues, its associated frequency occurrences are used to determine the different articles containing them. Then review these articles to determine the correct relations of these cues and add them to their correct relations.

All verification phases are executed iteratively, in which statistical reports are updated dynamically until reaching a satisfied Arabic rhetorical relations hypothesis.

5. Rhetorical Representation

The abstract structure of most texts is a tree. Most articles and text theories mention explicitly or implicitly that trees are good mathematical abstractions [6]. This descriptive framework provides a rhetorical representation by building a binary tree that depends on the relationships between the paragraphs that are defined within the matching relation cues phase, in order to distinguish between paragraphs in terms of a *satellite* or *nucleus*. As in Figure 4, if a paragraph P_N starts with any of the defined cue phrases, then there is a direct relation with the previous paragraph P_{N-1} and this paragraph P_N is then called a *satellite*. If the paragraph P_N starts without any of the defined cue phrases, then there is a relation with the first paragraph (root), at this situation, the paragraph P_N is called a *nucleus*. After the units are identified and the rhetorical relations between units are recognized, text rhetorical structures can be represented by a tree called an *RS-tree*.

Experiments show the satellite paragraphs percentage is more than 80% and nucleus paragraphs percentage is less than 20% as in figure5.

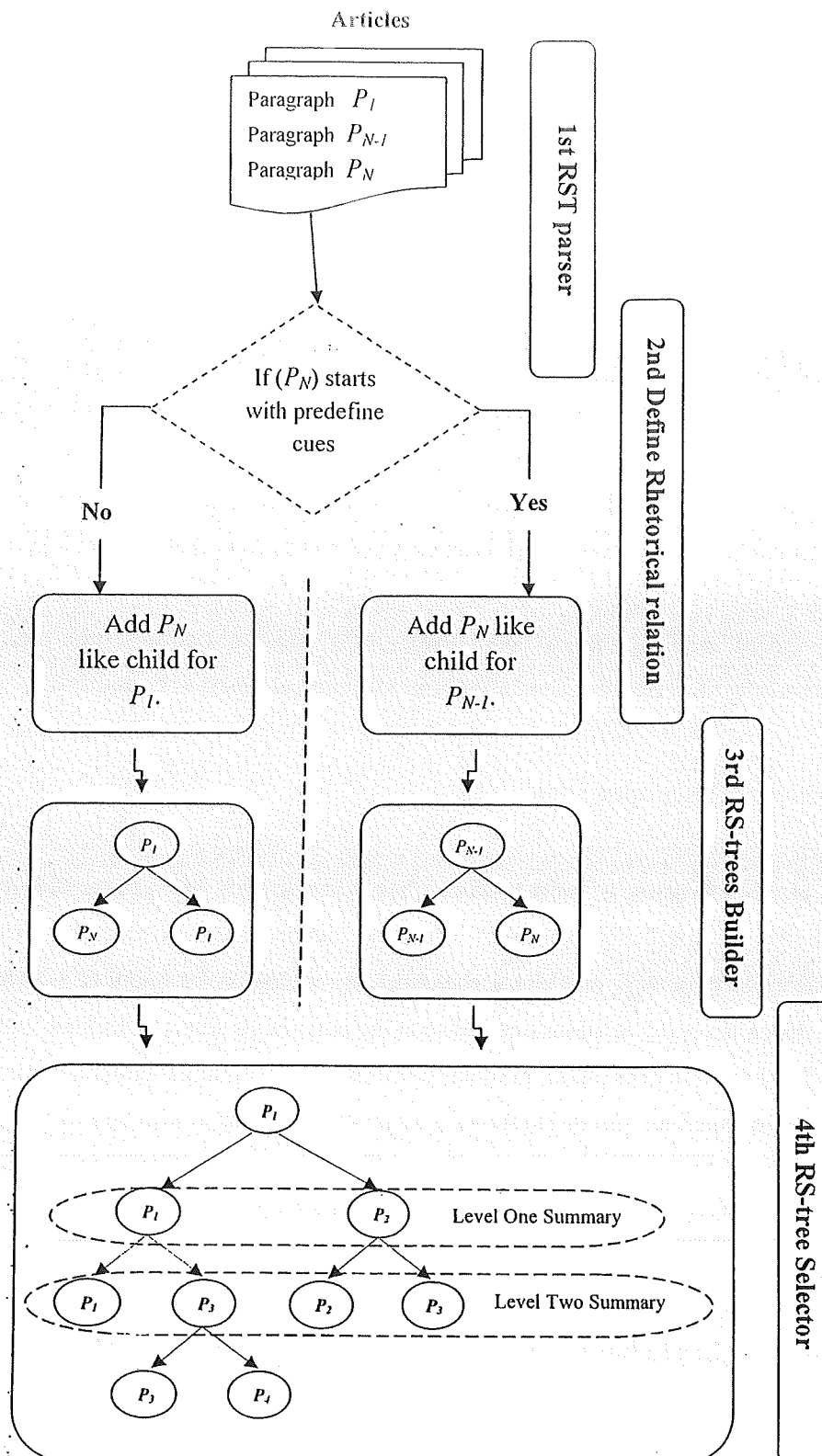


Figure 4: purposed rhetorical representation framework.

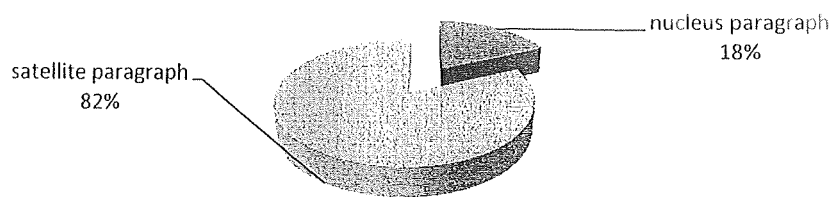


Figure 5: frequent ratio between nucleus to satellite paragraphs at corpus

6. Conclusion

This descriptive framework can analyze discourse structure and provides statistical results show that relations identified correctly within this framework, in order to be able to define the nucleus and satellite paragraphs. Overall results show nucleus paragraphs representing 20% that represent more important parts of text, and the satellite paragraphs represent 80% from all paragraphs at corpus.

7. References

- [1] Mann, William C., & Thompson, Sandra A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87-190). Marina del Rey, CA: Information Sciences Institute.
- [2] Mann, William C., & Thompson, Sandra A. (1988). *Rhetorical Structure Theory: Toward a functional theory of text organization*. *Text*, 8(3), 243-281.
- [3] Taboada, Maite. & Mann, William C. (2006a). Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4), 567-588.
- [4] Taboada, Maite, & Mann, William C. (2006b). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3), 423-459.
- [5] Marcu, D. (1997), "The Rhetorical Parsing Summarization, and Generation of Natural Language Texts", PhD Thesis, Department of Computer Science, University of Toronto.
- [6] Marcu, D. (2000) *The theory and practice of discourse parsing and summarization*, The MIT press.
- [7] Marcu, D. (2000) *The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach*. *Computational Linguistics*, To appear.

- [8] Hamida, M. (1997) *Nzam alartbat wa alrabi fi trkip algomla alarabia*. Dar Nubar for printing, ISBN 977-16-0235-7.
- [9] Alsana, W., Tourir, A. and Mathkour, H. (2005) Towards a Rhetorical Parsing of Arabic Text, The International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC05).
- [10] Hovy, E. and C-Y. Lin. (1999) .Automated text summarization in summarist. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. MIT Press.
- [11] Marcu, D. (1999). Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, chapter 11, pages 123–136. Cambridge, MA: The MIT Press.

201

202

203

204



جامعة القاهرة



معهد الدراسات والبحوث الإحصائية
المؤتمر السنوي السادس والأربعون
للإحصاء وعلوم الحاسب
وبحوث العمليات



علوم الحاسب والمعلومات

