Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge

تصميم وتنفيذ نظام مبني على المعرفة لأسترجاع النص بالأعتماد على سياق النص والمعرفة المسبقة للمستخدم

By

Ghaith AbdulSattar Alkubaisi

400920272

Supervisor

Dr. Hussein H. Owaied

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Master Degree in

Computer Science

Faculty of Information Technology

Middle East University

Amman, 2013

Authorization Statement

I, Ghaith AbdulSattar Alkubaisi, authorize the Middle East University to supply a copy of my thesis, Paper and electronically to libraries, organizations, or bodies and institutions concerned with research and scientific studies upon request.

Name : Ghaith Alkubaisi

Date : 19/8/2013

Signature : .....

اقرار تفويض

انا غيث عبدالستار الكبيسي افوض جامعة الشرق الاوسط بتزويد نسخ من رسالتي ورقيا والكترونيا للمكتبات، او المنظمات،او الهيئات و المؤسسات المعنية بالأبحاث والدراسات العلمية عند طلبها.

الاسم : غيث عبدالستار الكبيسي

التاريخ : ١٩/٥/٢٠١٢م

التوقيع : .....

Examination Committee Decision

This is certifying that the thesis entitled "Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge" was successfully defended and approved on August 19th 2013.

Examination Committee Members **Signature**

Dr. Hussein H. Owaied
Associate Professor, Department of Computer Science
(Middle East University)

Dr. Mudafar M. Al-Jarrah
Assistant Professor, Department of Computer Science
(Middle East University)

Dr. Basem M. F. Al Hadidy
Associate Professor, Department of Computer Information System
(Al-Balqa Applied University)

قرار لجنة المناقشة

نوقشت هذه الرسالة وعنوانها :- " Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge "

وأجيزت بتاريخ 19 / 8 / 2013

اعضاء لجنة المناقشة		التوقيع
1. الدكتور : حسين عويد الشمري	رئيسا ومشرفا :	
2. الدكتور : مظفر منير الجراح	عضوا :	
3. الدكتور : باسم محمد فاضل الحديدي	عضواً خارجيا (جامعة البلقاء التطبيقية)	

DECLARATION

I do my hereby declare the present research has been carried out by me under the supervision of Dr.Hussaein H. Owaied, and this work has not been submitted elsewhere for any other degree, fellowship or any other similar title.

Signature: 

Date: 19/8/2013

Ghaith AbdulSattar Alkubaisi

Department of Computer Science

Faculty of Information Technology

Middle East University

ACKNOWLEDGEMENTS

Initially thanks and appreciation to my supervisor, Dr. Hussein Owaied, who was the teacher and directed that did not spared the time or effort or knowledge from my studies Bachelors and through post-graduate, and thanks to all my colleague at work, Dr. Yasser Sadek, who lend a helping hand to me in the most important stages preparatory to the thesis, and thanks non uninterrupted for guys long road my father and mother, who lend a helping hand to me to connect to that stage interval of my life, and extend my sincere thanks to fellow Derby my wife, and also thank Information Technology Faculty members at the Middle East University for Graduate Studies, and my friends who worked on the encouragement me to complete this stage.

DEDICATION

Dedicate my humble efforts to my country (Iraq), and to my father and my mother, my wife, my kids, my brother and sisters and all my friends And praise be to Allah, Lord of the Worlds.

Table of Contents

List of Figures

Chapter One

Chapter Two

Chapter Three

Chapter Four

List of Tables

Chapter One

Chapter Four

List of Bar Graphs

Chapter Four

List of Abbreviations

AI.....Artificial Intelligence

API.....Application Programming Interface

CNF.....Conjunctive Normal Form

CPU.....Central Processing Unit

DB.....Data Base

DNF.....Disjunctive Normal Form

FK.....Foreign Key

GIF.....Graphic Interchange Format

HTML.....Hypertext Markup language

HTTP.....Hyper Text Transfer Protocol

ID.....Identifier

IR.....Information Retrieval

IRS.....Information Retrieval System

IRSs.....Information Retrieval Systems

IRT.....Information Retrieval Technology

JPEG..... Joint Photographic Experts Group

KB.....Knowledge Base

KBS.....Knowledge Based System

KID.....Knowledge Identifier

KKw.....Knowledge Keyword

KKwID.....Knowledge Keyword Identifier

Kw.....Keyword

Kws.....Keywords

ODP.....Open Directory Project

PCs.....Personal Computers

PK.....Prior Knowledge

PKID.....Prior Knowledge Identifier

PNG..... Portable Network Graphics

SE.....Search Engine

SQL.....Structured Query language

UP.....User Profile

UPK……User Prior Knowledge

WWW.....World Wide Web

XML..... Extensible Markup language

Abstract

The motive of this thesis is the desire to facilitate Information Retrieval Technology based on user profile and context, by using approaches and operators used in the field of Information Retrieval technology.

This thesis presents the design and implementation of a knowledge-based system for text retrieval. The system based on specific knowledge in the field of search technique and access to all information on the World Wide Web (www). The specific knowledge in this system as its knowledge base lead to the possibility of achieving the linking process between the searching texts entered by the user into the system and his interest areas together with the key words. The system consists of five parts; these are user interface, knowledge-base, inference engine, user profile as a database, and the search engine.

The proposed approaches for retrieving information from the World Wide Web through the most popular search engines GOOGLE and YAHOO, the proposed approaches work to retrieve information by creating a database that represents the user profile which contains a set of keywords to represent user areas interest, connects these keywords with search keyword or keywords in the search text box by using an appropriate Boolean operator. That is lead to generate context to help search engines to retrieve useful information from the World Wide Web.

The methodology used in this thesis is analytical and empirical methodology, analytical because it based on litterateur scurvies deal with IR technology and depend statistical studies on non-random sampling from society interested with this topic.

This study empirical study because need to implement an application to achieve the proposed approach

Usually when there are many key words for searching about particular information the search technique will be slow since there are many relationships between each key word with the other two concepts which are user prior knowledge and user profile.

The proposed system, to implement this proposed approach, is a windows application uses Visual BASIC.NET 2010, to design user interface, In addition to the use of SQL Server 2008 software to build a database store each user profile.

The proposed system can be as a facility available in the search engines in order to help naive users. The proposed method facilitates the user for retrieving information just within their areas of interest, reducing the number of sites that appear to the user on web browser and reducing the number of keywords entered by the user in the search text box.

Keywords : Boolean Operators, Context, Information Retrieval Technology, Knowledge Based System, Knowledge Representation, Search Engine, Searching Techniques, User Profile, Web Information Retrieval, World Wide Web.

الملخص

كان دافع هذه الرسالة هو الرغبة في تسهيل تكنولوجيا استرجاع المعلومات بالاعتماد على ملف المستخدم وسياق التص، وذلك من خلال استخدام طرق وعمليات مستخدمة في مجال تكنولوجيا ارجاع المعلومات.

يقدم هذا البحث تصميم وتنفيذ نظام مبني على المعرفة لاسترجاع النص. نظام يقوم على معرفة محددة في مجال تقنية البحث والوصول إلى جميع المعلومات على شبكة الانترنت. المعرفة المحددة في هذا النظام تتمثل في إمكانية تحقيق عملية الربط بين النص الذي تم ادخاله من قبل المستخدم في النظام واهتمامات المستخدم بالأضافة الى مجموعة من الكلمات الرئيسية. يتكون النظام من خمسة أجزاء هي واجهة المستخدم، قاعدة المعرفة، محرك الاستدلال، ملف المستخدم كقاعدة بيانات، ومحرك البحث.

ان الطريقة المقترحة تستخدم لاسترجاع المعلومات من الشبكة العالمية من خلال محركات البحث الاشهر استخداماً (GOOGLE & YAHOO)، ان الطريقة المقترحة تعمل على استرجاع المعلومات عن طريق بناء قاعدة بيانات تمثل ملف المستخدم تحتوي على مجموعة من الكلمات التي تمثل مجالات اهتمام المستخدم ، يتم ربطها مع الكلمة او الكلمات المدخلة في مربع النص بأستخدام عمليات منطقية حتى تولد سياق نص يساعد محرك البحث على الوصول للمعلومات الافضل على الشبكة العالمية.

المنهجية المستخدمة في هذه الرسالة هي تحليلية عملية, تحليلية وذلك لانها تعتمد على دراسات سابقة ضمن مجال ارجاع المعلومات, ودراسة احصائية اجريت على عينة غير عشوائية من المجتمع تهتم في هذا المجال. عملية لانها تعتمد على تنفيذ تطبيق يحقق الهدف من الطريقة المقترحة.

وهنا لابد من ان نذكر ان زيادة عدد الكلمات التي تمثل المفاتيح الاساسية في عملية البحث من الممكن ان يؤدي الى عدم الوصول الى البيانات المطلوبة بشكل جيد ومطابق لما يطمح اليه المستخدم.

ان النظام المقترح لتطبيق الطريقة المقترحة يمثل تطبيق من تطبيقات الويندوز وتم استخدام بيئة الفجوال بيسك دوت نت 2010 لتصميم واجهة مستخدم سهلة الاستخدام, بالاضافة الى استخدام برمجية SQL Server 2008 لبناء قاعدة بيانات يخزن عليها كل مايتعلق بالمستخدم لهذا النظام وتسمى بملف المستخدم.

هذا النظام صمم ليساعد الاشخاص الذين ليس لديهم المعرفة بتقنيات البحث وطرق استرجاع المعلومات للوصول الى المعلومات التي تطابق ما يبحثون عنه في الشبكة العالملية من خلال محركات البحث, ان الطريقة المقترحة تسهل على المستخدم استرجاع المعلومات ضمن مجالات اهتمامه فقط , وذلك من خلال تقليل عدد المواقع التي تظهر للمستخدم بالاضافة الى تقليل عدد الكلمات التي يدخلها المستخدم في مربع البحث.

Chapter One

Introduction

1.1 Overview

The World Wide Web (WWW) became the main source of information for anybody in the world. The huge amount of information on the web makes reach specific information difficult process, and there is only one way to find or get the information by using a Search Engine (SE). (Meredith et al. 2010)

For search engine users are difficult to get the specific information directly because of some reasons like Keywords (Kws) entered in searching text box on SE interface. By using the suitable Kws and a search technique approach from the user side can reach user to their specific information directly without any effort. (Kazunari et al. 2004)

If a SE that caters to all users, despite differences in their interests (areas of interest) for example, there is the user wants to search for articles, reports and researches to the scientist in the field of medicine, but it is possible there will be a scientist has the same name but in the field of computer science, so that will determine the desired domain is the area of interest of the user that sets it apart from the search engine through its dependence on ways (approaches) to help the user to achieve a useful information.

One of these proposed approaches, approach consisting of two processes, the first one build a user profile contains the Kws represent the user areas of interest, The second process is linking the cows that the user entered in the search text box in the

SE interface with Kws in the user profile for the user to retrieve information related to user areas of interest. (Kazunari et al. 2004)

The process of building user profile represents the knowledge acquired for the SE added by the user for identifying user area of interest to SE, help user to achieve specific information on the web.

To achieve the proposed approach it must be propose the design and implementation of the system help user who is searching on the WWW, and allow user to build a special profile contain Kws related with other areas of interest linked with entered Kws by user on search text in the SE interface. So this system must be provided with knowledge in the field of search techniques at WWW and the process of building a database representing user profile within the same system.

The proposed system based on specific knowledge and their role is the mediator between the user and the SE even easier for the user to retrieve information related to the user area of interest. This system is called Knowledge-Based System (KBS), which typically consists of three parts, user interface, inference engine, and the knowledge base. (Owaied, 2012)

1.2 Knowledge Based System

In the beginning of the second half of the twentieth century increased the world's need to build systems their goals are helping human in his decision making during daily works, with less time and effort. These systems are called expert systems, were created in the 1970s then increased in the 1980s in the medical domain such as MYCIN, ADVISER, and EMYCIN. The expert system is a special type of Knowledge Based System (KBS) and can be defined as a Knowledge Based System

with the facilities of correctly deduction and correctly decision making together with the editing facilities. (James, 2010)

1.2.1 Knowledge Based System Structure

The most important part of the knowledge-based system and also the expert systems is the knowledge base, Knowledge Based System = (knowledge + problem solving methods).

Therefore to build the knowledge-based system will be started from the knowledge base and then proposing the inference engine and user interface according to the forms used for representing the knowledge base (Ajith, 2005), (Owaied, 2012). The general structure of Knowledge Based System is shown in figure 1.1.

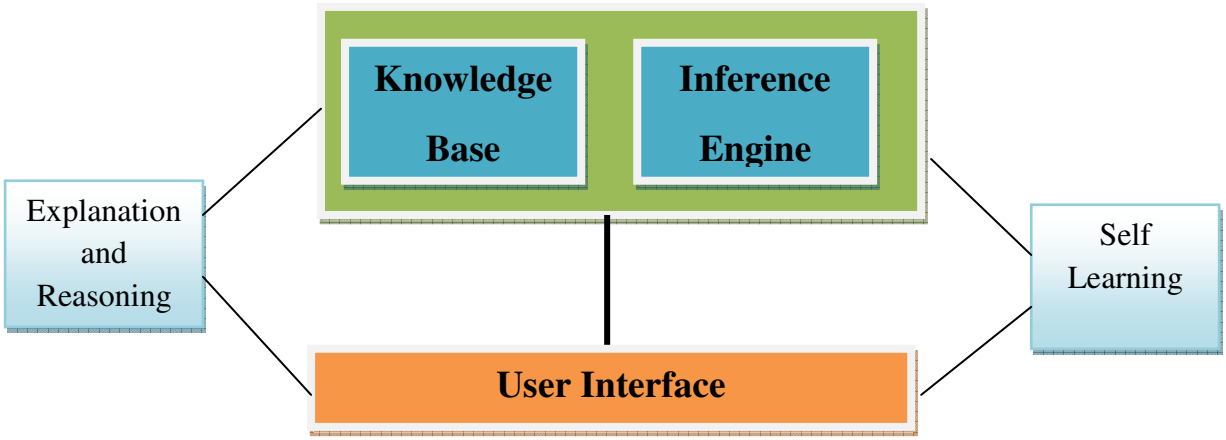


Figure (1.1) KBS Components (Priti, 2010)

Inference Engine is a computer program that tries to derive answers from a knowledge base, Inference algorithm, or computer implementation of it in a programming language, that allows carrying out inference steps to derive knowledge from knowledge automatically. If this knowledge is represented in the form of standard logic, then this term is synonymous with deduction (Priti & Akerkar, 2010).

Knowledge base captures the domain specific knowledge, and an inference engine that consists of algorithms for manipulating the knowledge represented in the knowledge base to solve a problem presented to the system. User Interface friendly interface to users working in their native language. (Cornelius, 1998)

1.2.2 Knowledge Representation

Since the forms used to represent the knowledge base affecting the proposing of the inference engine and user interface will introduce the most used forms of human for knowledge representation, such as rule base, frame base, semantic net base, case base, model base, etc. In the next subsection brief descriptions of the forms have been used for design knowledge-based systems. (Natasha, 2012)

The problem of knowledge representation occurs at four levels (Dagobert, 1989):

1. The general approach to knowledge representation.
2. The conceptual schema defining the nature of the data in the database through specifying entity types and relationship types.
3. Lists of entity values for each entity type.
4. The actual data and knowledge.

1.2.2.1 Rule base

The rule base is the set of rules which represents the knowledge about the domain. The decisions about how to process those data are almost invariably made by logic hard coded in the language of the program and stored in memory during program execution.

Knowledge Based System programmed by using a set of rules. These rules direct the computer to perform certain actions depending upon which rule is applicable to the current state of the program. (M Sasikumar et al. 2007)

The rule base is the conversion of human knowledge into the computer system by translating that knowledge into a set of rules, rule base has been applied in a vast number of application areas, for example about rules If-then rules, for example : {Rule 1: If A and C then Y, Rule 2: If A and X then Z, Rule 3: If B then X, Rule 4: If Z then D}. Rule base played an important role in modern intelligent systems and their applications in strategic goal setting, planning, design, scheduling, fault monitoring, diagnosis and so on. (Ajith, 2005)

A Rule-based system has four components set of rules, working memory, set of assertions that denote facts in some domain, rule applies, problem solving procedure, matching, control strategy and Conflict resolution when more than one rule can be applied. An example of a rule in the MYCIN Knowledge Based System (S. Quaglini et al. 2001)

Example of rile in the MYCIN Knowledge Based System :

If: (1) the strain of the organism is gram-positive.

(2) The morphology of the organism is coccus.

(3) The growth conformation of the organism is clumps.

Then there is suggestive evidence (0.7) that the identity of the organism is staphylococcus.

A typical rule based system consists of three components, they are working memory, rule base and inference engine shown in the figure1. 2.

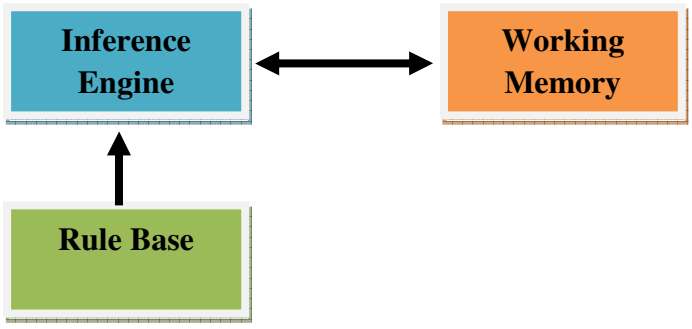


Figure (1.2) Rule Base components (M Sasikumar et al. 2007)

The Working Memory (WM) represents the set of facts known about the domain, such that user profile in the proposed system for this thesis.

1.2.2.2 Frame base

The frame is defined as a data structure with typical knowledge about a particular object or concept, were first proposed by Marvin Minsky in the 1970s. Frames provide a natural way for the structured and concise representation of knowledge. In general frame is an application of object-oriented programming for expert systems, also Expert Systems are required not only to store the knowledge but also to validate and manipulate this knowledge, so we need methods to add actions to our forums. (Anglin et al. 2004)

1.2.2.3 Semantic nets

Is a simple representation scheme that uses a graph of labeled nodes and labeled, directed arcs to encode knowledge, semantic nets is a directed graph which consists of nodes and links, the node presents an object and link presents the relationship between objects. The semantic nets are an alternative technique as a form of knowledge representation, The idea is that we can store our knowledge in the form of a graph, with nodes representing objects in the world, and arcs representing

relationships between those objects, The analysis in semantic nets has the location of an action represented as a node which is linked to the node showing the action arcs, for example, figure 1.3 presents an example of semantic net representation.

Figure (1.3) Example about semantic nets (William & James, 2005)

1.2.2.4 Case base

Case base is a methodology for problem solving that focuses on the utilization of past experience. It is based on solutions, information and knowledge available in similar problems previously solved, The implementation of this method requires the existence of a knowledge base that contains the cases that contain previous experience, note that cases has been collected from expert agents (Abdel-Badeeh et al. 2005)

1.2.2.5 Model base

Base model is an alternative technique as represented knowledge by using functional and behavioral description, to explain system operations and components and all relationships between their components using digital system design such that smart draw system (Eric 2004).

1.3 Information Retrieval System

Traditionally information would appear in journals or company reports, but increasingly it can be found online at the WWW. Tools to support information access and discovery on the Internet are proliferating at an astonishing rate. Some of this development reflects real progress but there are also many exaggerated claims. (Hadeel, 2009)

The goal of information retrieval is to find all documents relevant for a user query in a collection of documents. Decades of research in information retrieval were successful in developing and refining techniques that are solely Word-based (Monika, 2000).

The Information Retrieval System (IRS) An information retrieval system is a software program that stores and manages information on documents, often textual documents but possibly multimedia. The system assists users in finding the information they need. It does not explicitly return information or answer questions. Instead, it informs on the existence and location of documents that might contain the desired information. IRS consists of three parts, query, matching rule, and information store, figure below the present IRS structure. (Djoerd, 2009)

There is a lot of definition for Information Retrieval (IR) technology depending on the area of web access, CS and Information Retrieval System (IRS) :

- Information Retrieval (IR) is also used to facilitate specific searches such as finding information where related to users areas of interest, the field of IR also covers supporting users in browsing or filtering information. Information Retrieval System (IRS) can also be special system by using an appropriate approach for information retrieval. In web search, the IRS has to provide

search over billion web applications stored on millions of computers represents WWW; the goal of the IRS is to provide users with this information that will satisfy their information need. (Lancaster and Warner 1993)

- Information Retrieval (IR) is a technology has been central to the success of the Web. Web based indexing and search systems such as Google and Yahoo have profoundly changed the way we access information. For the semantic web technologies have an impact, they will have to be compatible with Web search engines and information retrieval technology in general. (Tim Finin et al. 2005)

1.3.1 Dimensions of Information Retrieval

IR is more than just text, and more than just web search although these are central people doing IR work with different media, different types of search applications, and different tasks. New applications increasingly involve new media such that video, photos, music, speech like text, content is difficult to describe and compare text may be used to represent them, IR approaches to search and evaluation are appropriate, Table 1.1 concluded all IR dimensions.

Table (1.1) Dimensions of IR (Susannah 2008)

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

1.3.2 Information Retrieval System Structure

The Information Retrieval System structure shown in figure 1.4

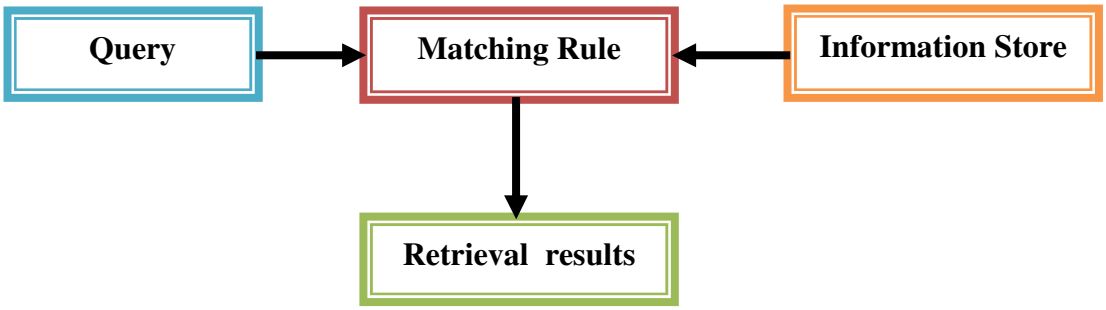


Figure (1.4) Information Retrieval System structure (Michael K. et al. 1994)

Query in IRS depend automatic classification of web queries in the SE, the topical classification of user queries allows for increased effectiveness, efficiency, and revenue potential in general-purpose web search systems. Such classification becomes critical if the system is to return results not just from a general web collection but from topic specific back end databases as well. Successful query classification poses a challenging problem, as web queries are very short, typically providing few features, this feature sparseness, coupled with the dynamic nature of the query stream and the constantly changing vocabulary of the average user hinders traditional methods of text classification.

Understanding the topical sense of user queries is a problem at the heart of web search. Successfully mapping incoming general user queries into topical categories, particularly those for which the search engine has domain specific knowledge, can bring improvements in both the efficiency and the effectiveness of general web search Much of the potential for these improvements exists because many of today's search engines, both for the Web and for enterprises, often incorporate the use of topic specific backend databases when performing a general web search. (Yogendra & Sandeep, 2011)

In general there are three kinds of IR applications depend on queries. (Yogendra & Sandeep, 2011):

1. Query recommendation: focusing on the reformulation of the original query, these kind of applications aims at identifying relationships between the original queries and alternative queries, such as generalization/specialization relationships.
2. Document recommendation: These kinds of applications will identify relevant documents to the original query.
3. Query classification: These kinds of applications will identify relevant queries and documents for each node in the directory, enriching their descriptions.

Matching Rule is an algorithm for matching between rule base and information on WWW, A number of retrieval models have been devised to abstract the processes underlying Information Retrieval systems. Models in which formal queries specify precise criteria for retrieved documents are said to be exact-match models, whereas best-match models return a ranked list of documents for a query conveying suitable documents. Exact-match models such as the Boolean model in which queries are formulated as logic expressions are more popular in legal and scientific search systems than Web search engines. (Clive, 2006)

The procedure of IR from WWW shown below in figure 1.5

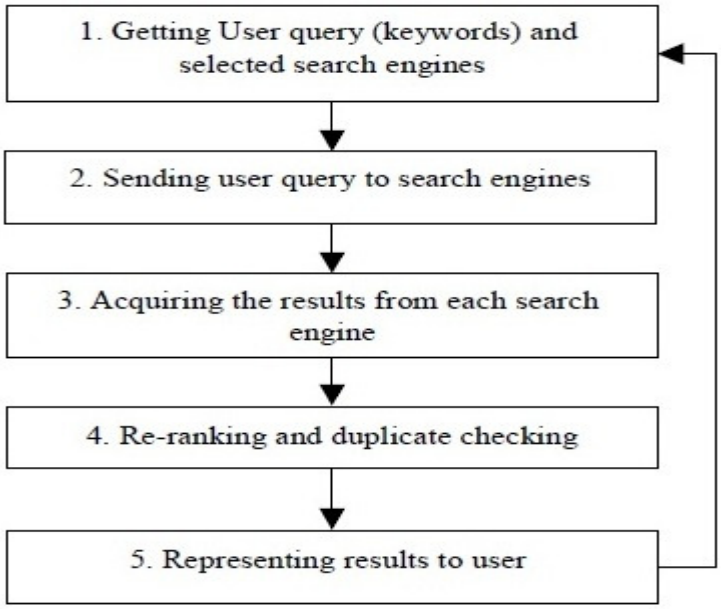


Figure (1.5) Work Flow of IR from WWW (Barfourosh & Anderson, 2002)

1.3.3 Main Problems Facing Information Retrieval System

1. Assisting the user in clarifying and analyzing the problem and determining information needs. Such assistance involves the following steps which describe the challenge is to design and implement systems that can provide such assistance. (Dagobert Soergel 1989) :

- Clarifying and analyzing the problem.
- Determining what part of the problem solution can be effected by the system and what part is left to the user.
- Determining what knowledge the user requires for her part in the problem solution.
- Determining what the user knows already.
- Deduce what information is necessary to lead the user from her present knowledge state to the required knowledge state.

2. Knowing how people use and process information, Involves in the following steps:
    - Relationship of information as to the problem-solving/decision-making process
    - How do people make relevant decisions.
    - How do people organize information in their minds, acquire it, process it for output.
3. Knowledge representation, a knowledge, information and database can be seen as an assembly of facts and rules such that the following example the first one about facts representation and the second about rule base representation.
    - ❖ Sample facts
        - MEDLINE is intended to inform physicians
        - For the purpose of patient care
        - Document-325 is about or relevant for Pneumonia
        - Penicillin cures Pneumonia
        - John has Pneumonia
    - ❖ Sample rule
        - IF person X has disease Y AND
        - Drug Z cures disease Y
        - THEN person X should-take drug Z
4. The human computer interface.

An information system user needs assistance in developing a query formulation that will produce a helpful information package in accordance with the specifications discussed earlier. Providing such assistance is the major function of the

user system interface. The interface must assist the user with problem clarification and with expressing the query in terms of the system and making best use of available system features.

The ideal information system interface supports two modes of interaction. In mode 1 the information system initiates a dialog without any initial user input (through displaying a menu or a fill-in-the-blank form). In mode 2 the system interprets an initial natural language query statement, transforms it into a first query formulation, and uses that query formulation as a starting point for a dialog. (Kamlesh, 2012)

Finally concluded that a Problem 1 and 2 deal with the user, her problems, information needs, and information processing behavior. Problem 3 and 4 deal with the knowledge representation and user interface.

1.4 Search Engine

The SE reflects the relationship between KBS and IR approach, It is fair to say that Internet based IR would collapse if search engines were not available, without search engines searchers would be about as successful negotiating the Internet as someone trying to look up a phone number in an unsorted phone book. While word of mouth pointers to pages from friends, acquaintances, and others are very useful, and the live hypertext links on the Web make it such a rich and convenient source of information, these means of negotiating the Internet do nothing for the user who does not even know where to begin looking that is the job of search engines. (Michael & Praveen, 1998)

The search strategy is essential if you hope to obtain satisfactory results. Most search engines index every word of a document, this method increase the number of search results retrieved while decreasing the relevance of these results. Most engines allow you to type in a few words, and then search for occurrences of these words in their database. Each one has their own way of deciding what to do about approximate spellings, plural variations, and truncation. (J. Sarkar, 2001)

Many search engines accept logic expressions of search terms. The user may specify keywords like AND, OR or NOT to denote Boolean operators (describe in details in section 1.5 in this thesis). In addition, many search engines offer modalities like to include or exclude of search terms. For example, the search term tunneling +IP - electrons indicate a search on the term tunneling which should return documents that not only contain the term tunneling but should also contain the term IP and should not contain the term electrons. The above query returned documents from GOOGLE out of which all of the results in the first few pages pertained to IP tunneling (Srinath Srinivasa & Bhatt, 2002).

Search engines provide three chief facilities, (1) They gather together a set of web pages that form the universe from which a searcher can retrieve information. (2) They represent the pages. (3) They allow searchers to issue queries, and they employ information retrieval algorithms that attempt to find for them the most relevant pages from this universe.

1.4.1 Types of Web Users

Web users can be broadly divided into three kinds based on their search strategies. These are (Srinath & Bhatt, 2002) :

a) Naive users, those users who do not have knowledge in the field of research over the internet.

b) Casual user searching the web for something that is loosely defined, The casual user searches the web for general information to satisfy his/her curiosity. The operations of the casual user are in the form of browsing the web starting from some arbitrary location, or in the form of queries over a web search engine. Usually for any keyword search, the user is likely to obtain vast amounts of matches. The less carefully the user's query is defined, the more likely are the chances of the user getting inundated with information. For obtaining the desired results it is important even for the casual user to work at precisely formulating the query.

c) A researcher looking for serious research level content over the web, The researcher typically looks to the web for information to help in his/her research. The search is more serious than that of a casual user, and is often augmented by other activities like annotations, bookmarking, etc. at the user's end. Search provided by the researcher is usually more complicated than a simple keyword search. It is necessary for the researcher to be much more precise in formulating search terms.

d) Professional looking for business intelligence by searching the web, business user searching the web for business intelligence typically looks for answers to questions like the following:

   Who are all my competitors?

   What is the market potential for our new suite of products? etc.

   Such a user requires more than search results from a search engine. Information required by such a user has to be extracted using inference mechanisms from results of conventional web searches.

1.4.2 Search Engine Structure

The simple structure for any search engine depend on the flows of information and IR process, figure 1.6 showing the base structure for any SE**.**

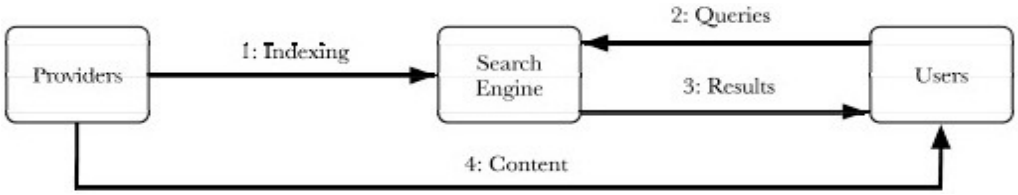


Figure (1.6) Information flow in SE (James, 2007)

The figure (1.7) shows the SE structure. For redundancy and fault tolerance, large search engines operate multiple, geographically distributed data centers. Within a data center, services are built up from clusters of commodity PCs. The type of PC in these clusters depends upon price, CPU speed, memory and disk size, heat output, reliability, and physical size. The total number of servers for the largest engines is now reported to be in the hundreds of thousands. Within a data center, clusters or individual servers can be dedicated to specialized functions, such as crawling, indexing, query processing, snippet generation, link-graph computations, result caching, and insertion of advertising content. (David, 2006)

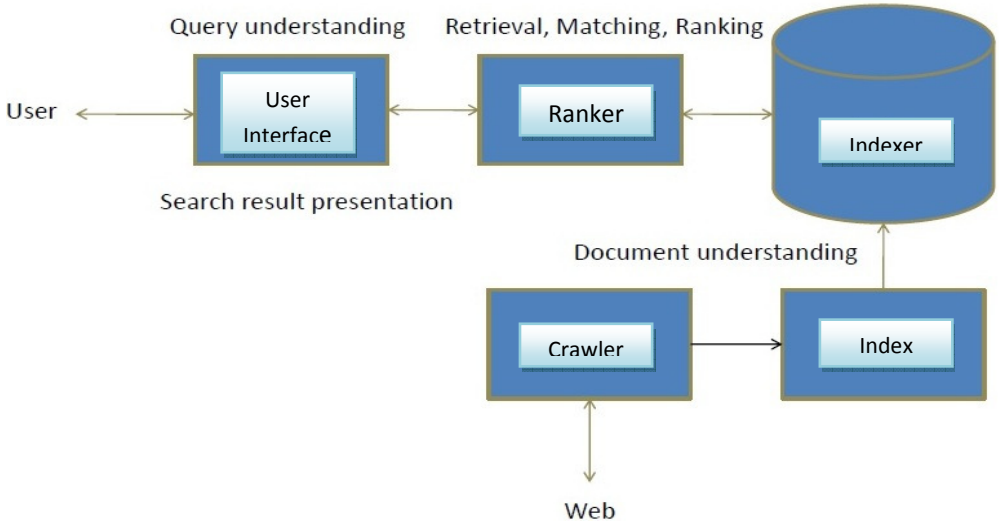


Figure (1.7) Search Engine Structure (Hang Li, 2011)

The crawler is CRAWLING ALGORITHMS the simplest crawling algorithm uses a queue of Unified Resource Locators (URLs) yet to be visited and a fast mechanism for determining if it has already seen a URL. This requires huge data structures, a simple list of 20 billion URLs contains more than a terabyte of data, Crawling proceeds by making a Hyper Text Transfer Protocol (HTTP) request to fetch the page at the first URL in the queue. When the crawler fetches the page, it scans the contents for links to other URLs and adds each previously unseen URL to the queue. Finally, the crawler saves the page content for indexing. Crawling continues until the queue is empty. (David, 2006)

For example about URLs is a web address and about Crawling is Traversing the Web by recursively following links from a seed.

Three important processes in Ranker the first one is retrieval finding documents from inverted index, matching calculating the relevance score between query and document pair, ranking documents based on relevance scores, importance scores, etc. Figure 1.8 shows matching between Query (q in figure 1.8) and Document (d in figure 1.8)

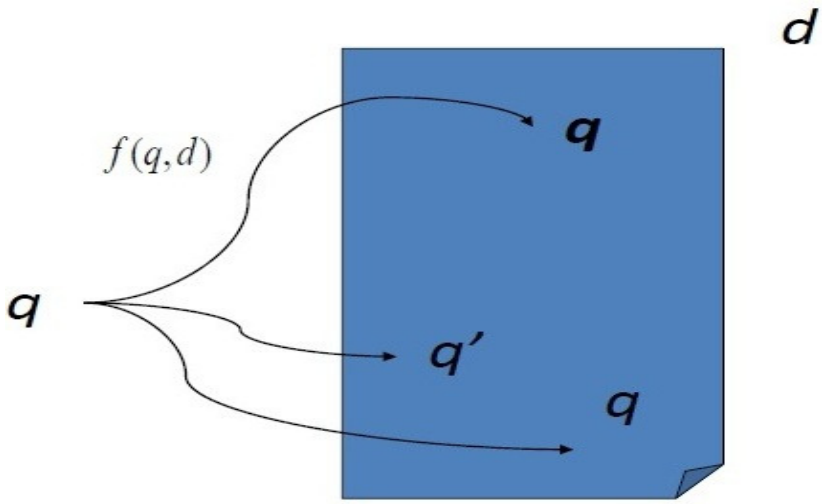


Figure (1.8) Matching between Query and Document (Hang Li, 2011)

1.4.3 Information Retrieval System and Search Engine

A search engine is the practical application of information retrieval techniques to large scale text collections; figure 1.9 displays the relationship between IRS and SE.

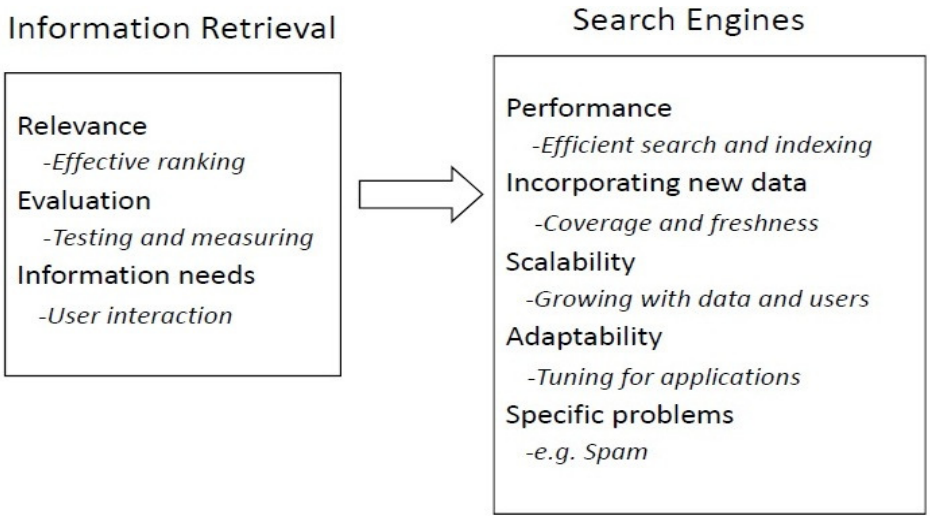


Figure (1.9) IRS and SE (Addison, 2008)

1.5 Rules for World Wide Web Information Retrieval

An online information seeker often fails to find what is wanted because the Keywords used in the request are different from the Keywords used in the relevant information. However, the searcher usually spends a significant amount of time reading retrieved information in order to determine whether it contains the information sought. (William et al. 2000)

The search engine uses an algorithm to achieve user request, this algorithm is kind of like atomic research. The search engine algorithm is a black box that you cannot see inside of, so you can only formulate theories and rules about a specific algorithm by testing its behavior. This is why it is so important to know all of the

different search engine operators (Rules for World Wide Web Information Retrieval) and how they work.

The new advance search leads to if they are not finding what they are searching for after using our basic search tips, try a search operator. Add one of these symbols (Boolean operators) to your search terms in the search text box of search engine interface directly to gain more control over the results that appear on the web browser. (Eileen, 2004)

1.5.1 Boolean Search Operators

Boolean operators are really useful for investigating how your competition uses keywords on the web. The basic Boolean search operators are AND, OR and NOT:

1. AND or (+) appear results that contain both of the keywords on the page. Since all of the major engines automatically search for all of the words you enter into the search field, the AND operator is usually unnecessary. If you want to find pages that contain keywords as a phrase, you should put quotation marks around your keywords ("keyword1 keyword2").

   The (+) operator tells the search engine to retrieve only documents that include the term it is used with. The (+) operator must be used before each term that must be included. For example curriculum (+) English (+) science (+) art will return documents that have all the words curriculum, English, science, art. Google and other search engines automatically add the AND operator between search terms.

This makes the (+) operator unnecessary. Most search engines treat (+) and AND the same way. (Lora & Dennis, 2006)

2. OR appear results that contain at least one of the keywords on the page. The OR operator is good to use when you are searching for more than one keyword and searching without an operator appear no results.
3. NOT or (–) appear results that contain one keyword but exclude the other keyword. This operator is especially useful to see which competitors are not using specific keyword combinations which you can then target.

Boolean operators help to refine your search terms to get better results by enabling you to expand, narrow, or focus your query.

Many search engines have included AND & NOT on their lists of stop words, or words that are excluded from your search. We recommend using (+), (-) , OR and quotation marks to investigate specific keyword combinations in GOOGLE and YAHOO search engines.

Otherwise the quote operator (" ") use to search for an exact word or set of words. This option is handy when searching for song lyrics or a line from literature. The quote operator turns two or more individual terms into a single phrase that is searched for together and in order. The quote operator can also be used to tell some search engines to include words that they may normally exclude, such as a, and, the or other common terms. (Lora & Dennis, 2006)

Internet search engines and many online databases such as journal indexes also use keywords and Boolean searching to help you tailor your search. By using Boolean operators (described in table 1.2 below) you may search for more than one term at a

time, specify logical relationships between terms, and define their proximity to one another. (Terre, 1999)

The rule which distinguishes the proposed system is the use of Boolean operator set directly in the search text box without any user effort to learn how and why , this reduces effort not only for users but for search engine to find appropriate algorithm help SE to retrieve the information which user need. Table 1.2, show a summary of Boolean operator.

Table (1.2) Boolean Operators (Margaret, 2010)

Operator	**Description**
And / +	Both operators do the same thing. Requires all terms to appear somewhere in the document, in any order Example : + curriculm+high+school
Not / -	Add a dash (-) before a word or set to exclude all results that include that word. This is especially useful for synonyms like Jaguar the car brand and jaguar the animal. Example : jaguar speed – car
OR	If you want to search for pages that may have just one of several words, include OR (capitalized) between the words. Without the OR, your results would typically show only pages that match both terms. Example : Olympics location 2014 OR 2018

1.6 Problem Definition

The purpose of this thesis is the design and implementation of Knowledge Based System for Information Retrieval from WWW, information related to users prior knowledge (areas of interest), and from the literature surveys does not find system retrieve information based on context and users prior knowledge with design rules based on Boolean operators to link Keywords directly on the search text box. Design this system facing the following problem:

1. Representation user's prior knowledge (areas of interest) in Data Base (DB).
2. Separation between users because each user has special areas of interest.
3. Identify rules for choosing a suitable Boolean operator to connect context with user profile.

1.7 Objectives

The main goal for this thesis is the design and implementation Knowledge Based System for retrieving useful information related to use prior knowledge, without any effort from the user. To achieve this goal need to satisfy the following:

1. Define the set of rules will be used in the knowledge base.
2. Convert the previous rules to computer system such as design an algorithm includes all these rules.
3. Reduce the number of links appear to use on the web browser.
4. Reduce the number of keywords that users enter in the search text.

1.8 Methodology

The methodology used in this study is analytical and empirical methodology, analytical because it based on litterateur scurvies deal with IR technology and depend statistical studies on non-random sampling from society interested in this topic. This study empirical study because need to implement an application to achieve the proposed approach.

Before started this work depend the questionnaire to collect data on a range of questions in the questioner see the Appendix, and the data are Numerical Data. The questionnaire distributed to Limited Community represented students in Al-Quds College; the number of samples is 100 students Distributors between the Department of Information Technology and Engineering Department and in different grades. This sample called simple random sample from a stratified sample class that targets a certain Limited Community. The questions and results represented in Questionnaire by using Simple Statistical Table operation added in chapter four.

1.9 Thesis Contribution

The technology of IR from the WWW which based on the user interested areas and their queries through the SE lead to create a special User Profile (UP) depending on log file by browsing information (history file) of user queries through the SE. This UP used from SE to retrieve related useful information corresponded user query but many users erase their log file to save their own privacy browsing information when other users use the same computer system. User maybe deleted his log file, this action lead to create new log file every time when user access SE, this process spends time to recognize UP. (Kazunari, 2004)

There are five basic approaches to user identification: software agents, logins, enhanced proxy servers, cookies, and session ids. Because they are transparent to the user, and provide cross-session tracking, cookies are widely used and effective. Of these techniques, cookies are the least invasive, requiring no actions on the parts of users. Therefore, these are the easiest and most widely employed. Better accuracy and consistency can be obtained with a login-based system to track users across sessions and between computers, if users can be convinced to register with the system and login each time they visit. A good compromise is to use cookies for current sessions and provide optional logins for users who choose to register with a site. (S. Gauch et al. 2007)

The proposed approach for IR in this thesis represented in creating DB includes Keywords represent other areas of interest or UP, these Keywords entered by user and they can update. These Keywords used later on for decreasing user effort in a number of entering Keywords in the search text box.

Create UP mean each user has a special account to give real user privacy, Which gives user lake of desire to erase any private data to him in the future, Contrary to what is customary in WWW IR applications. Linked UP Keywords with user text query by using Boolean operators directly on search text box, this process facilitates to achieve related information.

So UP include Keywords entered by user represent other areas of interest and linked process between UP keywords and Keyword or Keywords in the search text box to inference context help SE to retrieve useful information for users matching their UP not browsing history file.

1.10 Thesis Structure

- **Chapter One**: The Introduction. Include overview about the Knowledge-Based System, Information-Retrieval System, Search Engines, Rules for World Wide Web Information Retrieval and Boolean operator to explain the Problem Definition, Objectives, Methodology and thesis Contribution.
- **Chapter Two**: Is the literature surveys for the thesis, showing the related work and comparison between the related work and thesis contribution.
- **Chapter Three**: Explain the design of Knowledge Based System for Text Retrieval with explain system the system components and their algorithms.
- **Chapter Four**: Preview the implementation of the proposed system with system testing for each rule, and the questioner statistical analysis.
- **Chapter Five**: Contains the conclusion and the future work for the thesis.

Chapter Two

Literature survey

Many researchers have been talked about IR technology, depending on the user's areas of interest are, represented in DB called user's profile, or deduced through the user log file represent history file of user internet browsing, and we will be talking about the role of KBS to improved IRS. We'll talk about these related works briefly.

2.1 Related Work

When talk about related works that is mean explain and compare some points in different researches to describe the relationship between this thesis subject and domain with other researchers in the field of IR and using KBS to improve IRS.

2.1.1 Knowledge Based System in Information Retrieval

Hsinchun (1995) talked about the role of KBS and their approach to building IRS, he saw the way for creating computer systems with knowledge or Intelligence has long been the goal of researchers in Artificial Intelligence (AI). Many interesting KBS have been developed in the past few decades for such applications as medical diagnosis, engineering troubleshooting, and business decision making. Most of these systems have been developed based on the manual knowledge acquisition process a significant bottleneck for KBS development.

A recent approach to knowledge elicitation is referred to as Knowledge Mining or Knowledge Discovery. Grounded on various AI based machine learning

techniques, the approach is automatic and it acquires Knowledge or identifies patterns directly from examples or databases.

IR research has been advancing very quickly over the past few decades. Researchers have experimented with techniques ranging from probabilistic models and the vector space model of the knowledge based approach and the recent machine learning techniques. Significant insights regarding how to design more useful and Intelligent IRS have been gained.

2.1.2 IR based on list of Keywords

Alexander (1998) descriptive approach for retrieving personal quires depends on user personalization domain. Said the popular internet browsers such as Microsoft Internet Explorer or the Netscape Navigator allow for organizing bookmarks in a personalized manner. Internet Browser can be use personalization file to filtering and rating the most popular links for the user based on personalization file.

An initial user profile may be provided in the form of a list of keywords. The user behavior is tracked while he reads activities like scrolling, peeking at maximizing open articles in new windows or saving them to scrapbook probably mean a user is interested in that article. User profile represents no explicit information is given, but the article suggests the profile is a list of weighted keywords for comparing and matching easily.

2.1.3 Adaptive Web Search and IR Based on Context and Log File

Kazunari et al. (2004) their work to relate between user query and information at WWW, he explains the previous goal by this example " for the query “Java” some

users may be interested in documents dealing with the programming language, “Java” while other users may want documents related to “coffee” Therefore, Web search results should adapt to users with different information needs. Then they say that In order to predict such information needs, there are several approaches applying data mining techniques to extract usage patterns from Web logs.

However, the discovery of patterns of usage data by itself is not sufficient for performing the personalization tasks. Therefore, these techniques are not so appropriate for Web personalization. Another novel information system designed to realize such adaptive systems have been proposed that personalize information or provide more relevant information for users.

Kazunari Sugiyama et al. inference three types of Web search systems:

a) Systems using relevance feedback.

b) Systems in which users register their interest or demographic information.

c) Systems that recommend information based on user ratings.

The figure 2.1 below showing the proposed model for IR

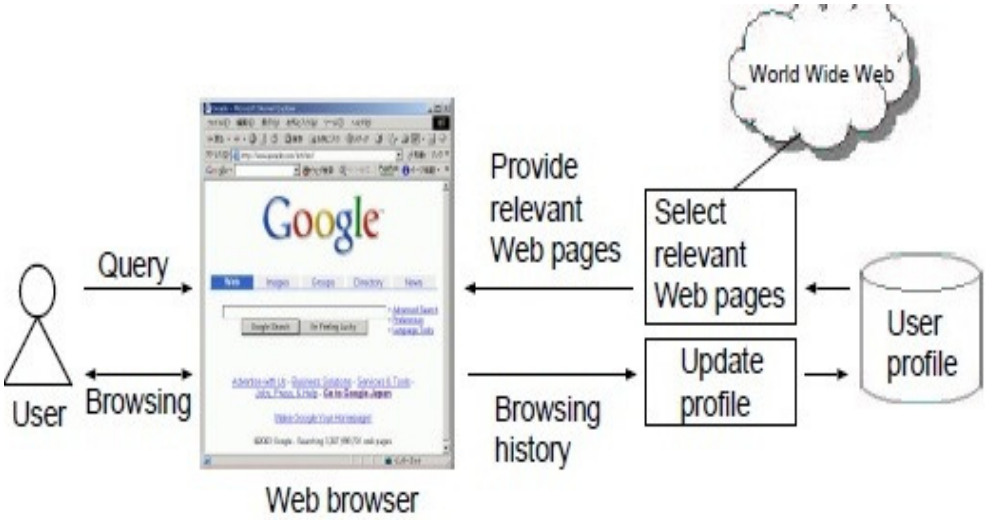


Figure (2.1) Proposed model for IRS (Kazunari et al. 2004)

Their proposed approaches to achieve IR personalization such there is :

a) Hyperlink Based: The field of Web IR focuses on hyperlink structures of the Web, for example with Web search engines such as Google to address several problems with these engines, (1) the weight of a Web page is merely defined, and (2) the relativity of contents among hyperlinked Web pages is not considered, they proposed several approaches to refining the scheme for Web pages using their hyperlinked neighboring pages. In personalized Web searches, the hyperlink structures of the Web are also becoming important. The use of the personalized Page Rank to enable personalized Web searches was first proposed in, where it was suggested as a modification of the global Page Rank algorithm, which computes a universal notion of the importance of a Web page.

b) Personalized Web Sites : Link topology and the structure and the contents of Web pages are often used in the construction of a personalized Web site.

c) Recommender System : As one of the most promising approaches to alleviate this overload, recommender systems have emerged in domains such as Ecommerce, digital libraries, and knowledge management. These systems provide personalized suggestions based on user preferences. Recommender systems collect user feedback in the form of ratings for items in a given domain and exploit similarities and differences among profiles of several users in determining how to recommend an item. There are two prevalent approaches to constructing recommender systems, collaborative filtering based and content based recommendation.

Finally they concluded user profile build depend on three approaches (1) relevance feedback and implicit approaches, (2) user profiles based on pure browsing history, and (3) user profiles based on the modified collaborative filtering.

B. Van et al. (2004) present a novel architecture for Information Retrieval on the Web called (Vimes). This architecture is based on a broader definition of relevance. This broader definition lies in the fact that there is more than just topical relevance. Documents (or: resources) must also conform to other constraints with regard to form, format and also things like price and quality.

They recognized that information retrieval systems can be personalized for users by means of profiles. During the last few decades a lot of research has been invested in the area of user profiles. Often, these profiles are used to enhance the query by capturing the user's notions of query terms. However, profiles can be used more extensively.

Example for explaining user profile approach for IR from WWW:

(User Profile) A (user) profile consists of a set of preferences with regard to the behavior of a search engine as well as constraints on the results it presents to the user. To illustrate this definition, the following list are the items that make up a particular user-profile:

- **Preferences:** prefer a maximum of 25 results per page, and by selecting a relevant resource (clicking on the link) will open a new window.
- **Constraints:** prefer HTML and PDF formats and refuse the Microsoft DOC format. Furthermore, the size of the resource should not exceed 25Mb.

Using this definition, there are two areas in the retrieval process where profiles can be used. Firstly, they can be used for post-processing the results of the ranking process, figure 2.2 below showing the proposed WWW Information Retrieval System.

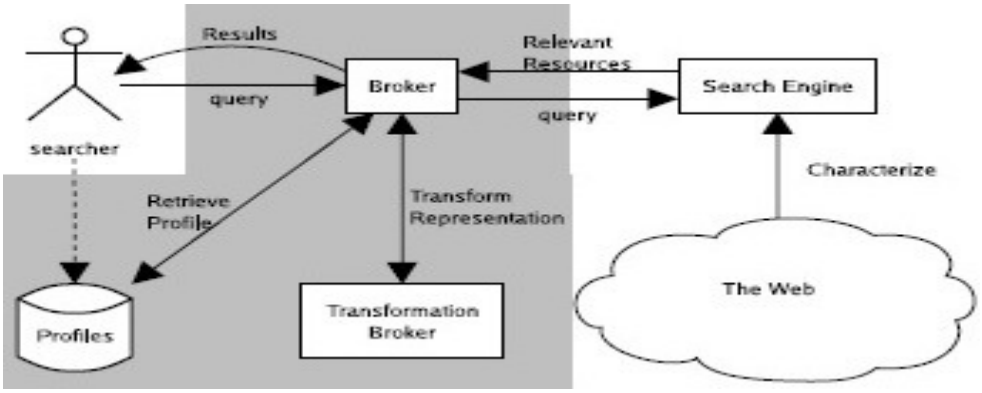


Figure (2.2) Proposed system for WWW information retrieval B. .van et al. 2004)

2.1.4 Personalized information retrieval based on context and ontological knowledge

Mylonas et al. (2004) focus on the combination of contextualization and personalization methods to improve the performance of personalized information retrieval. The key aspects in our proposed approach are the explicit distinction between historic user context and live user context lead combined in order to improve the accuracy and reliability of personalization for retrieval.

Historic mean The user's usage history comprises of a combination of all types of actions, provided that a user is able to perform any type of action at a given time. An association between the related history documents and concepts exists through the utilization of the semantic index, which is a priori constructed during analysis of either the raw content, or the associated textual annotation.

They said , The notion of context has been long acknowledged as being of key importance in a wide variety of fields, such as mobile and pervasive computing, computational linguistics automatic IR, the representation and usage of context as a key element e.g. to enhance the understanding of human speech, needs, activities and intentions, to raise the system awareness of the external conditions that may influence human priorities and plans, to build an awareness of the available resources for the

system to accomplish a certain goal, and in general, to better grasp the relative nature of truth.

Context is an increasingly common notion in IR, and has been identified as a major challenge in the field of IR. This is not surprising since it has been long acknowledged that the whole notion of relevance, at the core of IR, is strongly dependent on context - in fact it can hardly make sense out of it. Several authors in the IR field have explored approaches that are similar to ours in that they find indirect evidence of searcher interests by extracting implicit meanings in information objects manipulated by users in their retrieval tasks. A key differentiating aspect of our approach is the use of semantic concepts, rather than terms (for example : strings), for the representation of these contextual meanings, and the exploitation of explicit information attached to the concepts available in a knowledge base.

2.1.5 The WWW Information Retrieval Based on User Activities

S. Gauch et al. (2007) they claim that in the modern Web, as the amount of information available causes information overloading, the demand for personalized approaches for information access increases. Personalized systems address the overload problem by building, managing, and representing information customized for individual users. This customization may take the form of filtering out irrelevant information and or identifying additional information of likely interest for the user. Research into personalization is ongoing in the fields of information retrieval, artificial intelligence, and data mining, among others

They say that Early personalization research focused on personalized filtering and rating systems for e-mail, electronic newspapers, Usenet newsgroups and Web documents. More recently, personalization efforts have focused on improving navigation effectiveness by providing browsing assistants and adaptive Web sites.

In order to construct an individual user's profile, information may be collected explicitly, through direct user intervention, or implicitly, through agents that monitor user activity. Although profiles are typically built only from topics of interest to the user, some projects have explored including information about non-relevant topics in the profile. In these approaches, the system is able to use both kinds of topics to identify relevant documents and discard non-relevant documents at the same time.

They explain how to build a user profile depend on user web activities. Profiles that can be modified or augmented are considered dynamic, in contrast to static profiles that maintain the same information over time. Dynamic profiles that take time into consideration may differentiate between short-term and long-term interests. Short-term profiles represent the user's current interests whereas long-term profiles indicate interests that are not subject to frequent changes over time. For example, consider a musician who uses the Web for his/her daily research. " One day, he/she decides to go on vacation, and she uses the Web to look for hotels, airplane tickets, etc. "

User profile should reflect her music interests as long-term interests, and the vacation-related interests as short-term ones. Once the user returns from his/her vacation, he/she will resume his/her music-related research, and the vacation information in his/her profile should eventually be forgotten. Because they can change quickly as users change tasks, and less information is collected, short-term user's interests are generally harder to identify and manage than long-term interests. In general, the goal of user profiling is to collect information about the subjects in which a user is interested, and the length of time over which they have exhibited this interest, in order to improve the quality of information access and infer user's intentions. Figure 2.3, displays the user-profile based on personalization

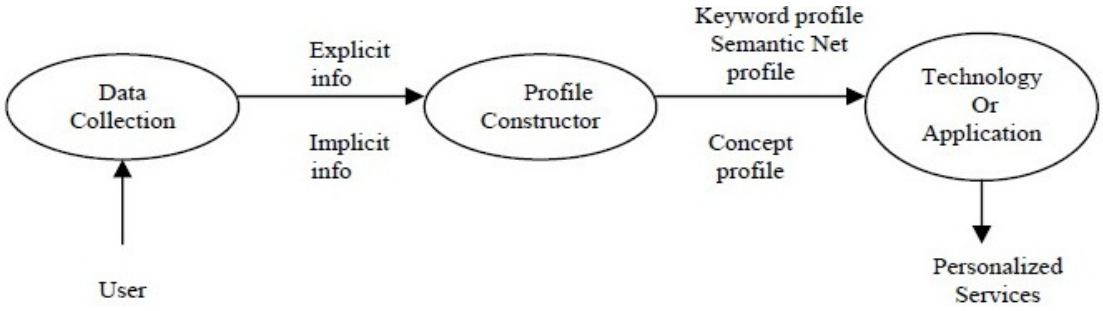


Figure (2.3) Building Personal User Profile (S. Gauch et al. 2007)

The most common representation for user profiles is sets of keywords. These can be automatically extracted from Web documents or directly provided by the user. Weights, which are usually associated with keywords, are numerical representations of users' interests. Each keyword can represent a topic of interest or keywords can be grouped in categories to reflect a more standard representation of a user's interests

The previous approach for building UP adopted in this thesis by giving weighted by number of rows in DB table represent UP.

Zhongmin et al. (2007) The Web provides an extremely large and dynamic source of information, and the continuous creation and updating of Web pages magnifies information overload on the Web. Both casual and non-casual users often use search engines to find a needle in this constantly growing haystack, who define a knowledge worker as someone has paid work involves significant time spent in gathering, finding, analyzing, creating, producing or archiving information, report that 59% of the tasks performed on the Web by a sample of knowledge workers fall into the categories of information gathering and finding, which require an active use of Web search engines.

Most existing Web search engines return a list of search results based on a user's query but ignore the user's specific interests and/or search context. Therefore, the identical query from different users or in different contexts will generate the same set of results displayed in the same way for all users, a so called one-size-fits-all.

Our proposed approach is a form of client-side personalization based on a Framework for area of interest and result categorization. It piggybacks on a standard search engine such as Google and categorizes and displays search results on the basis of known user interests. As a novel feature of our approach, the mapping framework automatically maps the known user interests onto a set of categories in a Web directory, such as the Open Directory Project (ODP) or YAHOO directory.

An advantage of this mapping framework is that, after user interests have been mapped onto the categories, a large amount of manually edited data under these categories is freely available to be used to build text classifiers that correspond to these user interests.

In summary, to generate user profiles for personalized search, previous studies have asked users for explicit feedback, such as ratings and preferences, or collected implicit feedback, such as search and browsing history. However, users are unwilling to provide explicit feedback even when they anticipate a long-run benefit. Implicit feedback has shown promising results for personalizing search using short-term context.

However, generating user profiles for long-term context through implicit feedback will take time and may raise privacy concerns. In addition, a user profile generated from implicit feedback may contain noise because the user preferences have been estimated from behaviors and not explicitly specified.

Myriam et al. (2009) they proposed approach personalizes data retrieval using implicit user information and interests measurements. As the data manipulated is expressed by attributes and values, we define several similarity measures. These measurements consider both semantic and spatial user contexts. The approach

personalizes Web content and especially spatial information focusing on its spatial semantic aspects.

The proposed approach personalizes data retrieval using implicit user information and interests measurements. We start, in the next section with related works presentation and discussion. We then present our architecture including user and data modeling approaches and the similarity measures used to increase the quality of the personalization process and the measures used to deduce user's interest

Our proposition is based on a dynamic and iterative construction of a multidimensional user model. This multidimensional approach is used to represent and describe the user towards different dimensions. The model proposed (noted Mu) is composed of 4 dimensions: user profile, spatial model, graphic model and textual

model. If we consider U as the set of users, a user u∈U will have as model M_u:

$$M_u = Pn \square Ds \square Dt \square Du$$

Where:

– Dt = keywords employed by the users for their textual search

– Ds = spatial positions of the users

– Dn = entities visited by the users

– P = user profile.

They concluded the Web personalization attracts rising research efforts to facilitate Web information retrieval and navigation. Generally, Web personalization

and user modeling approaches do not focus on the spatial aspect of the information and the constraints it implies. In this paper, we have presented some background knowledge on existing Web and spatial Web personalization systems.

Arthur et al. (2011) their proposed system that provides the user to register with it and based on the users registered areas of interest the system searches the related and efficient information from the world wide web using the technique of web text mining and arranges the unstructured data into structured format and present it to the user. This system also stores the previously searched data and based on users Areas of interest and rating awarded to the interest of the user his profile will be updated at a particular scheduled time.

The proposed system enables user to edit his profile and specify his area of interest and award rating to it. Our application provides the login support and lets him to edit his profile in which he will be specifying his area of interest. Our application scans the user profile and extracts the area of interest as the key word (this part of our application is referred to as Term Extractor. The key word is taken as input and the information related to that is searched first in the repository. If the information is found in the repository it will be organized in a proper format and dumped back to the profile.

The next time a user logs in to his account, he will find the information related to his area of interest which he has specified in his profile. If the information is not found in the repository then search will be carried out in the web and the information is gathered and redirected to the user in proper format. The searched information is also stored in our repository so that any second user who enters the same area of

interest can be served by searching in the repository itself rather than searching in the web again and again if.

Searching for the information in the system database and the internet, using the keywords extracted from the Extractor Search Engine searches the related information in the system database and if the search is successful then the user profile is provided with the retrieved information and if the search is not successful than the search engine performs the search over the web and retrieves the information from the web and update the system database and also the user profile.

2.1.6 The Role of Boolean Operator for Enhance IR

Venkat N.et al. (1997) they claim that The Boolean model represents documents by a set of index terms, each of which is viewed as a Boolean variable and valued as True if it is present in a document. No term weighting is allowed. Queries are specified as arbitrary Boolean expressions formed by linking terms through the standard logical operators: AND, OR, and NOT. Retrieval status value (RSV) is a measure of the query-document similarity. In the Boolean model, RSV equals 1 if the query expression evaluates to true, RSV is 0 otherwise. All documents whose RSV evaluates to 1 are considered relevant to the query.

And then explain the most important algorithm need Boolean operator these are WebCrawler has a robot that starts with a known set of HTML documents and uses the URLs in them to retrieve new documents. The search engine directs the navigation in a modified breadth-first mode. It maintains a list of Web servers and URLs to fetch from them, which it does in a round robin fashion to avoid fetching documents consecutively from the same server.

WebCrawler aims at indexing at least one document from each server. Users can also submit URLs. It indexes both the title and full text of HTML documents, and

its index is updated weekly. Terms are weighted by their frequency of occurrence in the document divided by their frequency in the reference domain. Terms that appear frequently in the document and infrequently in the reference domain are heavily weighted, while those that appear infrequently in either are given lower weights. WebCrawler supports full Boolean and phrase searches.

Robert & Manning (1998) they talk about using Boolean queries or ranking documents using document and term weights will result in better retrieval performance has been the subject of considerable discussion among document retrieval system users and researchers. We suggest a method that allows one to analytically compare the two approaches to retrieve and examine their relative merits. The performance of information retrieval systems may be determined either by using experimental simulation, or through the application of analytic techniques that directly estimate the retrieval performance, given values for query and database characteristics. Using these performance predicting techniques, sample performance figures are provided for queries using the Boolean And and OR, as well as for probabilistic systems assuming statistical term independence or term dependence.

Retrieval systems based on Boolean logic have long served as the cornerstone of the commercial document retrieval system market and remain very important because of the relative simplicity of the query language and the ease with which it can be understood and implemented. The most common use of a Boolean expression is to state what characteristics must be present in the material to be retrieved in a system that retrieves and presents to users bibliographic records or full-text. A second use of Boolean expressions likely to increase in importance over the next decade is in rules incorporated into the document and email filtering systems. Such a rule might take the form of a statement.

Boolean expressions typically use three operators: and, or, and not. A search for documents about both dogs and cats might be expressed as dogs and cats. Logical Implications, such as dog implies mammal, if something is a dog then it is a mammal, may be expressed without using the implication operator.

They conclude that, it becomes necessary to treat Boolean systems as special forms of probabilistic retrieval systems. We suggest a way to do this here, by comparing the ranking provided by individual Boolean operators with the ranking provided by systems consistent with probabilistic models. Any Boolean query may be expressed in either of the common normalized forms of Boolean expressions: Conjunctive Normal Form (CNF), or Disjunctive Normal Form (DNF). CNF represents the conjunction of disjunctions, that is, a series of "anding" components with these components, in turn, consists of the "oring" of individual terms (or the negations of these terms.) Any Boolean expression can be converted into CNF. Similarly, a logical expression in DNF is a disjunction of conjunctions, a set of "ored" components, where each component consists of anding terms.

By converting a Boolean expression of these normal forms, a ranking of documents using these probabilistic methods may be easily implemented through the simple combination of the methods for the Boolean or, not, and ands. Below we assume that all our queries have been converted to CNF, thus simplifying the types of operands each of our Boolean operators must accept.

Bernard & Caroline (2003) they said web searchers seldom use advanced query structure, such as Boolean operators or phrase searching, when using information retrieval (IR) systems. Numerous Web studies note the near absence of query operators such as AND, OR, NOT, MUST APPEAR (+), and PHRASE (" ") in

Web queries. It is generally assumed that the proper use of query operators would increase the effectiveness of Web searches.

All the search engines supported all the query operators in some form, but there are frequently minor changes to the searching rules. At the time of the study, America Online Search (AOL) directly supported the use of the AND, OR, MUST APPEAR, and PHRASE operators from its main page, although it also provided an advanced search option that facilitated the use of operational functionality. Google directly supported the AND, OR, MUST APPEAR, and PHRASE operators, although it states that the use of AND is not necessary. MSN directly supported the AND, OR, and MUST APPEAR operators. There was a drop down box for PHRASE searching. All search engines provided an advanced search mode, which directly supported all of the operators considered here as well as other features.

2.2 Summary

From the previous studies can be concluded the following points :

1. The IR technology fully depends on context which reflect the user request.
2. All IRS from WWW based on Logs File (History of Browsing) to represent UP.
3. UP represent the user areas of interest which inference from Logs File and URL Ranking process.
4. Using Boolean operators in link context (represents a user query) by the SE after user query by using special algorithms for Boolean operators.

Note of the foregoing that the characteristic of our study in the field of the development of IR technology from previous studies is :

1. UP represents by using DB include Keywords reflect the user areas of interest.
2. UP able to update by the user.
3. UP help user to reduce the number of Keywords entered in the search text box.
4. UP reflects areas of interest directly by user not by inference depend on user Logs File.
5. Context represents the relationship between user areas of interest and search text.
6. The relationship between user areas of interest and text create by using Boolean operators directly in the text box to reduce the effort for SE because here does not need to use special algorithms to achieve this relationship.

Chapter Three

System Architecture and Implementation

In this chapter we will explain the system proposed model shown in figure (3.1), and describe the way of IR through it. Then we explain each system components, relationships between components, algorithms and how the system work.

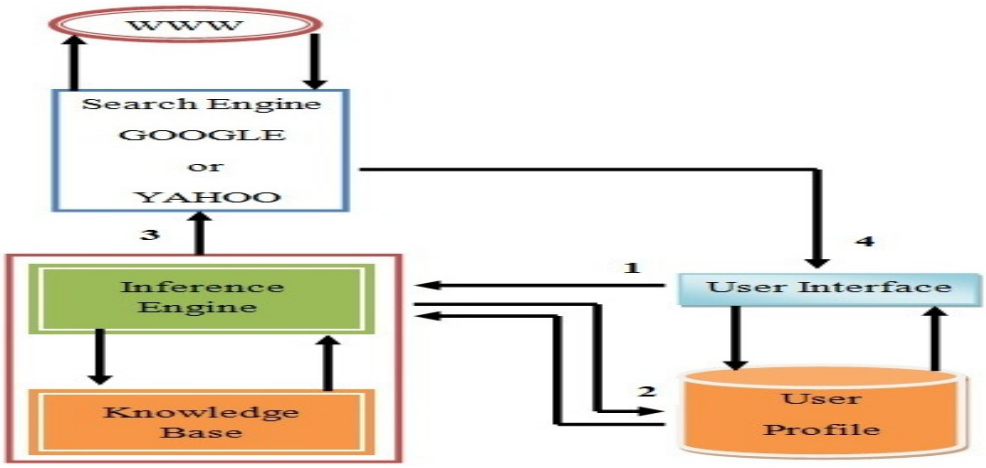


Figure (3.1) System Proposed Model

3.1 How System Works

The methodology used in this system through construction of the relationships between the three concepts, user searching text, user prior knowledge, and the set of key words. These relationships will be saved in the knowledge base as rules for future search lead to inference strongly context in the search text box. From the proposed model we have seen step numbers from 1 to 4 to explain the procedure of system work as follows:

1. The user sends queries from the user interface.

2. System inference context after select keyword from DB (User Profile).

3. Write context on the SE search text box.

4. Browsing useful information related to user areas of interest on User Interface.

3.2 System Components and Rules

In this section we will explain the four system components User Interface, User Profile, Knowledge Base and inference engine respectively to explain their role for IR from WWW.

3.2.1 User Interface

In this system the role of user interface is intermediate between the user and search engines (GOOGLE and YAHOO), User interface include many features to help user he/she to create his/her special profile help SE for retrieving information from WWW related with user profile and browsing this information as a link on user interface. User Interface includes four related windows each one have some features help user to execute IR process.

1. Log in window : given privacy and specialization to the user and make his/her profile independent.

2. Registration window : lead user to create his/her won profile

3. Main window : display the main two features user profile and searching process

4. Searching and Browsing window : user can be searched and browsing retrieved information.

3.2.2 User Profile

The User Profile (UP) is data base connected to the user interface, include keywords represents user areas of interest. This data base has two tables, the first one saves all keywords represents user areas of interest save as User prior Knowledge (UPK) and the second one represents a user account privacy such as user name and password to give independence for each user profile.

1. Table one (Users) : include seven records explain the most important
    a. User ID : help user to save his/ her privacy
    b. User password : to improve user privacy and independence
    c. Used Major : one keyword reflects the closed user area of interest
2. Table two (Prior Knowledge) : include for records
    a. Keyword ID : represents the weight of keyword record in a database table (KID).
    b. Knowledge Field : represent the closed user field or the main user area of interest.
    c. Knowledge Keywords (KKws) : keywords each one represents user area of interest.
    d. User ID : represent the Foreign key (FK) of Table one and two relationship .

3.2.3 Knowledge Base

Is a special kind of database for knowledge management. A knowledge base is an information repository that provides a means for information to be collected, organized, shared, searched and utilized. It can be either machine-readable or intended for human use Machine-readable knowledge base store knowledge in a computer-readable form, usually for the purpose of having automated deductive reasoning applied to them. They contain a set of data, often in the form of rules that describe the knowledge in a logically consistent manner. An ontology can define the structure of stored data. Logical operators, such as And (conjunction), Or (disjunction), material implication and negation may be used to build it up from simpler pieces of information. Consequently, classical deduction can be used to reason about the knowledge in the knowledge base. Some machine-readable knowledge bases are used in artificial intelligence, for example as part of an expert system that focuses on a domain like prescription drugs or customs law. Such knowledge bases are also used by the semantic web. (Priti & Rajendra, 2010)

From the previous definition KB represent a set of facts, events, processes or procedures, and meta-knowledge for the specific and narrows certain domain store in DB called knowledge base. In this system KB represented by set of Rules store in class to help the inference engine unit to inference the suitable relationship between UPK and UP to generate strong context write on SE search text box help SE matching rule for retrieving the best related information. We will explain these functions respectively:

3.2.3.1 Default Rule

This combination of rules represent the simple context generation when the user has one area of interest without any details related to this area.

a. Default Rule Pseudo Code Using (+) Operator

[1] PKID as integer; (ID = ID number of record in Table)

[2] KKwID as integer;

[3] n as integer ; (variable represents ID number)

[4] PK as char;

[5] KKw as char;

[6] [**Rule D.a1**] If (PKID = = 0 && KwID = = null) Then

[7] Search textbox = Search Kw + PK ;

[8] [**Rule D.a2**] Else If (PKID = = 0 && KwID = = 0) Then

[9] Search textbox = Search Kw + PK + + Kw ;

[10][**Rule D.a3**] Else

[11] Search textbox = Search Kw + PK + + KwID = = 0 + +.....+ + KwID = = n ;

[12] End If

b. Default Rule Pseudo Code Using (-) Operator

[1] PKID as integer; (ID = ID number of record in Table)

[2] KKwID as integer;

[3] n as integer ; (variable represents ID number)

[4] PK as char;

[5] KKw as char;

[6] [**Rule D.b1**] If (PKID = = 0 && KwID = = null) Then

[7] Search textbox = Search Kw - PK ;

[8] [**Rule D.b2**] Else If (PKID = = 0 && KwID = = 0) Then

[9] Search textbox = Search Kw + PK - Kw;

[10][**Rule D.b3**] Else

[11] Search textbox = Search Kw + PK - KwID = = 0 + +.....+ + KwID = = n ;

[12] End If

c. Default Rule Pseudo Code Using (OR) Operator

[1] PKID as integer; (ID = ID number of record in Table)

[2] KKwID as integer;

[3] n as integer ; (variable represents ID number)

[4] PK as char;

[5] KKw as char;

[6] [**Rule D.c1**] If (PKID = = 0 && KwID = = null) Then

[7] Search textbox = Search Kw **OR** PK ;

[8] [**Rule D.c2**] Else If (PKID = = 0 && KwID = = 0) Then

[9] Search textbox = Search Kw + PK **OR** Kw ;

[10][**Rule D.c3**] Else

[11] Search textbox = Search Kw + PK **OR** KwID = = 0 + +.....+ + KwID = = n ;

[12] End If

3.2.3.2 Intersection Rule

This combination of rules represent the complex context generation when user has more than one area of interest with Kws as a detail related to this area.

a. Intersection Rule Pseudo Code (many PK and many Kws) by Using (+ and OR) Operators

[1] PKID as integer ;

[2] KKwID as integer ;

[3] n as integer ;

[4] PK as char ;

[5] KKw as char ;

[6] [**Rule I.a1**] If (PKID > = 0 && Kw= = null) Then

[7] Search textbox = Search Kw + PKID = = 0 OR Search Kw + PKID > 0 ;

[8] [**Rule I.a2**] Else If (PKID > 0 && KwID = 0) Then

[9] Search textbox = Search Kw + PKID = = 0 + + KwID > = 0 OR Search Kw + PKID > 0 ;

[10] [**Rule I.a3**] Else If (PKID > 0 && KwID > PKID) Then

[11] Search textbox = Search Kw + PKID = = 0 + + KwID > = 0 OR Search Kw + PKID > 0 + + KwID > PKID ;

[12] [**Rule I.a4**] Else

[13] Search textbox = Search Kw + PKID = = 0 + + KwID = = 0 to < (PKID > 0) + + OR Search Kw + PKID > 0 + + KwID > (PKID > 0) ;

[14] End If

b. Intersection Rule Pseudo Code (many PK and many Kws) by Using (+ and -) Operators

[1] PKID as integer ;

[2] KKwID as integer ;

[3] n as integer ;

[4] PK as char ;

[5] KKw as char ;

[6] [**Rule I.b1**] If (PKID > = 0 && Kw= = null) Then

[7] Search textbox = Search Kw + PKID = = 0 - Search Kw + PKID > 0 ;

[8] [**Rule I.b2**] Else If (PKID > 0 && KwID = 0) Then

[9] Search textbox = Search Kw + PKID = = 0 + + KwID > = 0 - Search Kw + PKID > 0 ;

[10] [**Rule I.b3**] Else If (PKID > 0 && KwID > PKID) Then

[11] Search textbox = Search Kw + PKID = = 0 + + KwID > = 0 - Search Kw + PKID > 0 + + KwID > PKID ;

[12] [**Rule I.b4**] Else

[13] Search textbox = Search Kw + PKID = = 0 + + KwID = = 0 to < (PKID > 0) - Search Kw + PKID > 0 + + KwID > (PKID > 0) ;

[14] End If

This system represents the prior knowledge and the knowledge keywords description (KKw) by levels depend on the ID number of records, so when the user input the first PK that is mean take an ID number (0) and any keywords follow PK represent the (kW) description, if user enter new PK that is mean take an ID number larger than the first PK and their (kW) description, then any keywords follow the second PK take an ID number larger than the second PK.

- Example about ID number Levels

PK No.1 : Civil Engineering..... ID = 0

Kw No.1 : Traffic..... ID = 1

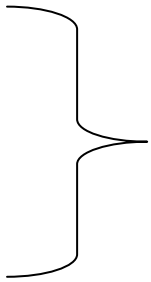
. . .

. . .

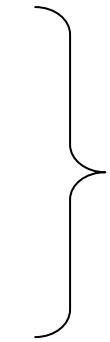
Kw No.n : Keyword .....ID = n

PK No.2 : Medicine.....ID > n

Kw No.x : Anatomy.....ID > PK No.2 ID



all keywords after PK No.1 just related with PK No.1



all keywords after PK No.2 just related with PK No.2

3.2.4 Inference Engine

Is a computer program that tries to derive answers from a knowledge base. It is the brain that expert systems use to reason about the information in the knowledge base for the ultimate purpose of formulating new conclusions. Inference engines are considered to be a special case of reasoning engines, which can use more general methods of reasoning. (Ajith, 2005)

3.2.4.1 Inference Engine Structure

The separation of inference engines as a distinct software component stems from the typical production system architecture. This architecture relies on a data store. The inference structure includes the following (Brian, 2005) :

1. The interpreter executes the chosen agenda items by applying the corresponding base rules.
2. The scheduler maintains control over the agenda by estimating the effects of applying inference rules in light of item priorities or other criteria on the agenda.
3. Consistency enforcer attempts to maintain a consistent representation of the emerging solution.

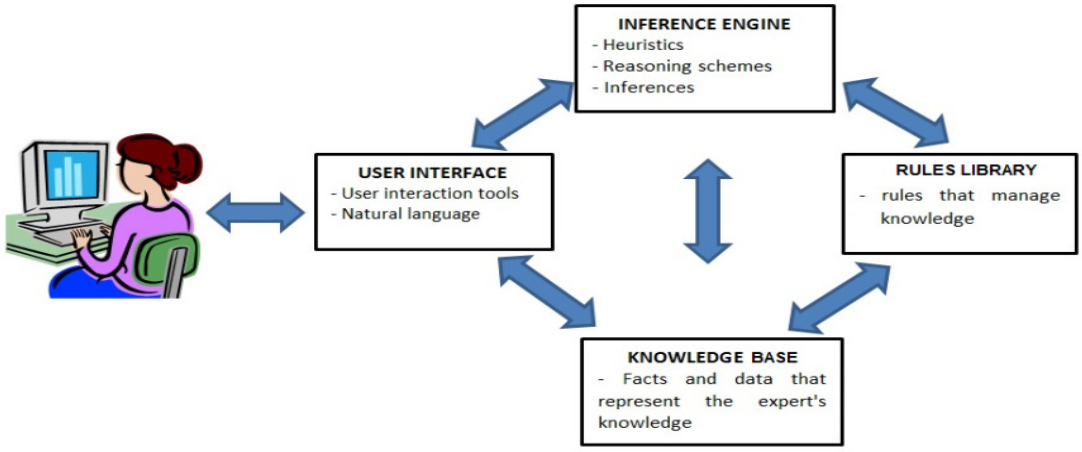


Figure (3.2) Inference Engine Work (Lucien et al. 2012)

The inference engine is an essential element of an expert system, since it works as the engine control that evaluates and applies the rules. In the process of problem-solving, these rules must be in accordance with the existing information in the working memory. (H. Araki 2005)

In this system inference engine unit work in inference strong context after choosing the appropriate rule form KB to use the suitable Boolean operator for connecting Kws.

3.2.4.2 Inference Engine Pseudo Code Rule

[1] If (PKID = = 0 or KwID > = 0) Then

[2] Call "GOOGLE API " or "YAHOO.COM" (User can retrieve information from each one)

[3] (Call Default Rule)

[4] Else

[5] Call "GOOGLE API " or "YAHOO.COM"

[6] (Call Intersection Rule)

[7] End If

3.2.4.3 System Algorithm of Calling GOOGLE API

```
Imports Google.API.Search
Public Class frmSearch

Dim htmlContent As String = ""
        WebBrowser1.Navigate("about:blank")
        WebBrowser1.Document.Write("")
        htmlContent = _
        "<head><meta http-equiv='Content-Type' content='text/html; charset=utf-
8'></head>" & _
        "<body>" & _
        "<table width=60% dir=ltr>" & _
        "<tr><td style='font-weight:bold;text-align:Left;Font-
size:16pt;'>Result Of : " & txtData.Text.Trim & "</td></tr>"
        Dim htmlMenuContent As String = ""
        Dim priorK As DataTable
        priorK = SearchRecordBySQL("SELECT * FROM priorKnowledge WHERE userID="
& userIDOf)
        Dim keyStr As String
        Dim AllkeyStr As String = ""
        If priorK.Rows.Count > 0 Then
            Dim i, j As Integer
            Dim VarprioKnoledge As String = ""
            Dim keys() As String

            For i = 0 To priorK.Rows.Count - 1
                keys =
priorK.Rows(i).Item("KnowledgeKeyWords").ToString.Split(";")
                keyStr = ""
                For j = 0 To keys.Length - 1
                    keyStr = keyStr & keys(j) & "+"
                Next

                AllkeyStr = AllkeyStr & "+" &
priorK.Rows(i).Item("KnowldegeField")
                If keyStr.Length > 1 Then
                    AllkeyStr = AllkeyStr & "+" & keyStr.Substring(0,
keyStr.Length - 1) & " OR "
                End If
            Next
            If AllkeyStr.Length > 0 Then
                AllkeyStr = AllkeyStr.Substring(0, AllkeyStr.Length - 3)
            End If
            MsgBox(AllkeyStr)
        End If
        Try
            Dim client As New GwebSearchClient("http://www.google.jo")

            Dim results As IList(Of IWebResult) =
client.Search(txtData.Text.Trim & " " & AllkeyStr, 1000)
            For Each result As IWebResult In results
                htmlMenuContent = htmlMenuContent & _
                "<tr><td style='text-align:left;Color:#00f;font-size:12pt;font-
weight:bold'><a href='" & result.Url & "'>" & result.Title & "</a></td></tr>" &
_
                "<tr><td style='text-align:left;Color:#0f0;font-size:10pt;'>" &
result.Url & "</td></tr>" & _
                "<tr><td style='text-align:left;Color:#000;font-size:11pt;'>" &
result.Content & "</td></tr>" & _
```

```
                "<tr><td style='text-align:left'> </td></tr>"
            Next
            htmlContent = htmlContent & htmlMenuContent & _
            "<td style='font-weight:bold;text-align:right'> </td></tr>" &
_
            "</table></body>"
            WebBrowser1.DocumentText = htmlContent
        Catch ex As Exception
            WebBrowser1.DocumentText = "<h1>Cannot Connect To The
Internet</h1>"
        End Try
    End Sub
```

3.2.4.4 System Algorithm of Calling YAHOO.COM

```
Dim htmlContent As String = ""
        WebBrowser1.Navigate("about:blank")
        WebBrowser1.Document.Write("")
        htmlContent = _
        "<head><meta http-equiv='Content-Type' content='text/html; charset=utf-
8'></head>" & _
        "<body>" & _
        "<table width=60% dir=ltr>" & _
        "<tr><td style='font-weight:bold;text-align:Left;Font-
size:16pt;'>Result Of : " & txtData.Text.Trim & "</td></tr>"
        Dim htmlMenuContent As String = ""
        Dim priorK As DataTable
        priorK = SearchRecordBySQL("SELECT * FROM priorKnowledge WHERE userID="
& userIDOf)
        Dim keyStr As String
        Dim AllkeyStr As String = ""
        If priorK.Rows.Count > 0 Then
            Dim i, j As Integer
            Dim VarprioKnoledge As String = ""
            Dim keys() As String

            For i = 0 To priorK.Rows.Count - 1
                keys =
priorK.Rows(i).Item("KnowledgeKeyWords").ToString.Split(";")
                keyStr = ""
                For j = 0 To keys.Length - 1
                    keyStr = keyStr & keys(j) & "+"
                Next

                AllkeyStr = AllkeyStr & "+" &
priorK.Rows(i).Item("KnowldegeField")
                If keyStr.Length > 1 Then
                    AllkeyStr = AllkeyStr & "+" & keyStr.Substring(0,
keyStr.Length - 1) & " OR "
                End If
            Next
            If AllkeyStr.Length > 0 Then
                AllkeyStr = AllkeyStr.Substring(0, AllkeyStr.Length - 3)
            End If
            '            MsgBox(AllkeyStr)
        End If
        Try
```

```
MsgBox("http://search.yahoo.com/search?p=" & txtData.Text.Trim & " " &
AllkeyStr & "&fr=yfp-t-721")
            WebBrowser1.Navigate("http://search.yahoo.com/search?p=" &
txtData.Text.Trim & " " & AllkeyStr & "&fr=yfp-t-721")

        Catch ex As Exception
            WebBrowser1.DocumentText = "<h1>Cannot Connect To The
Internet</h1>"
        End Try
    End Sub
```

Chapter Four

Testing and Validation

In this section we will display how the system takes input then retrieve information depend on the input keyword or keywords. The test represented by level each one depend on the required rule.

4.1 Test 1 : User enters one PK without any keyword

The procedure of this level start when users enter a keyword from his/her mind, just one keyword, connected with PK then inference context forward to SE to retrieve related information browsing on user interface. The following figures 4.1,4.2 and 4.3 preview IR result of this level depend on UP respectively.

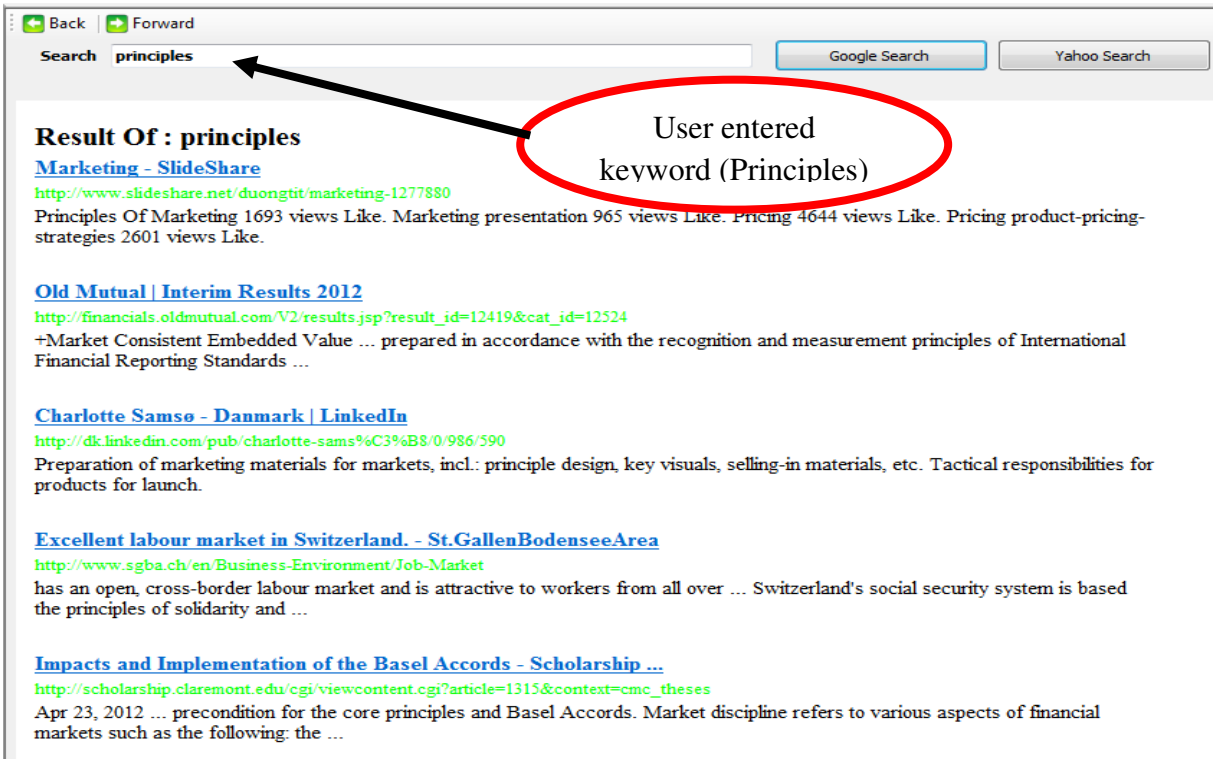


Figure (4.1) Results of Test1 Depend on GOOGLE API

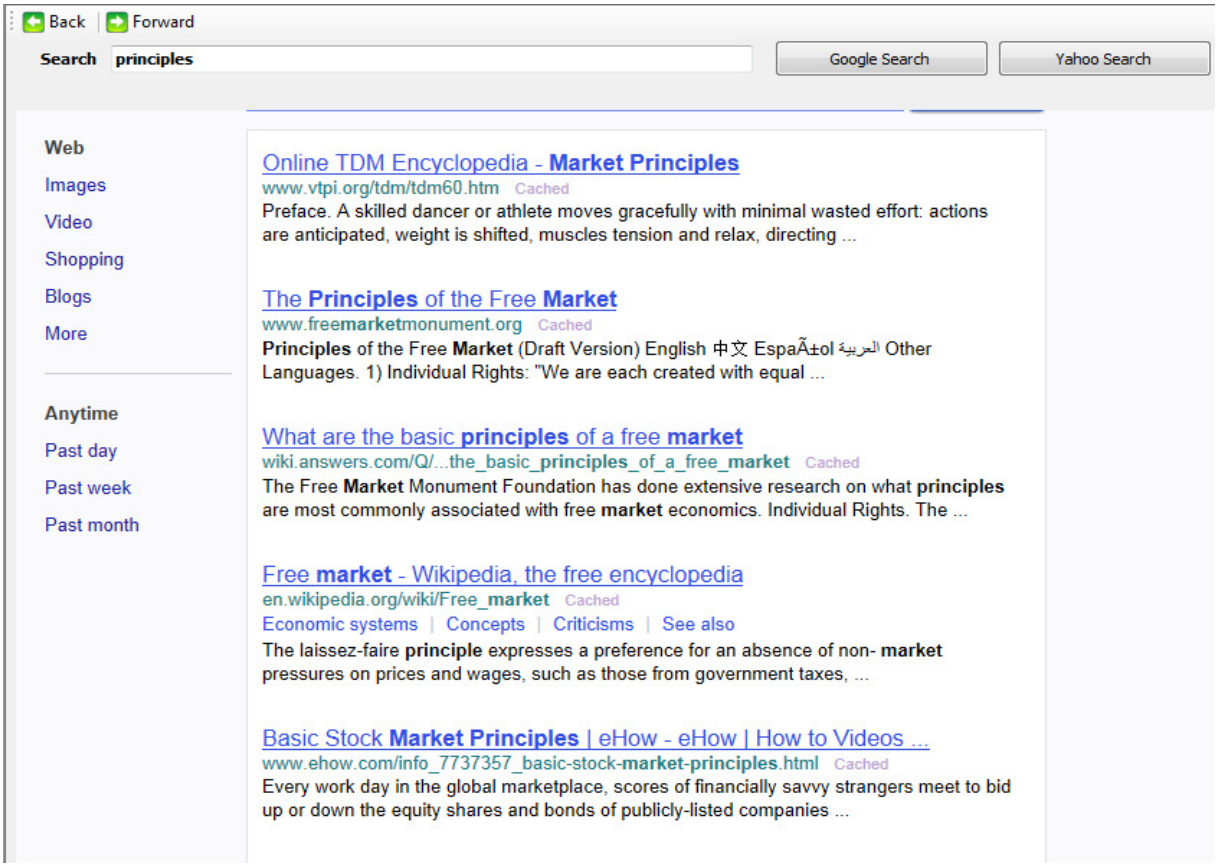


Figure (4.2) Results of Test1 depend on YAHOO.com

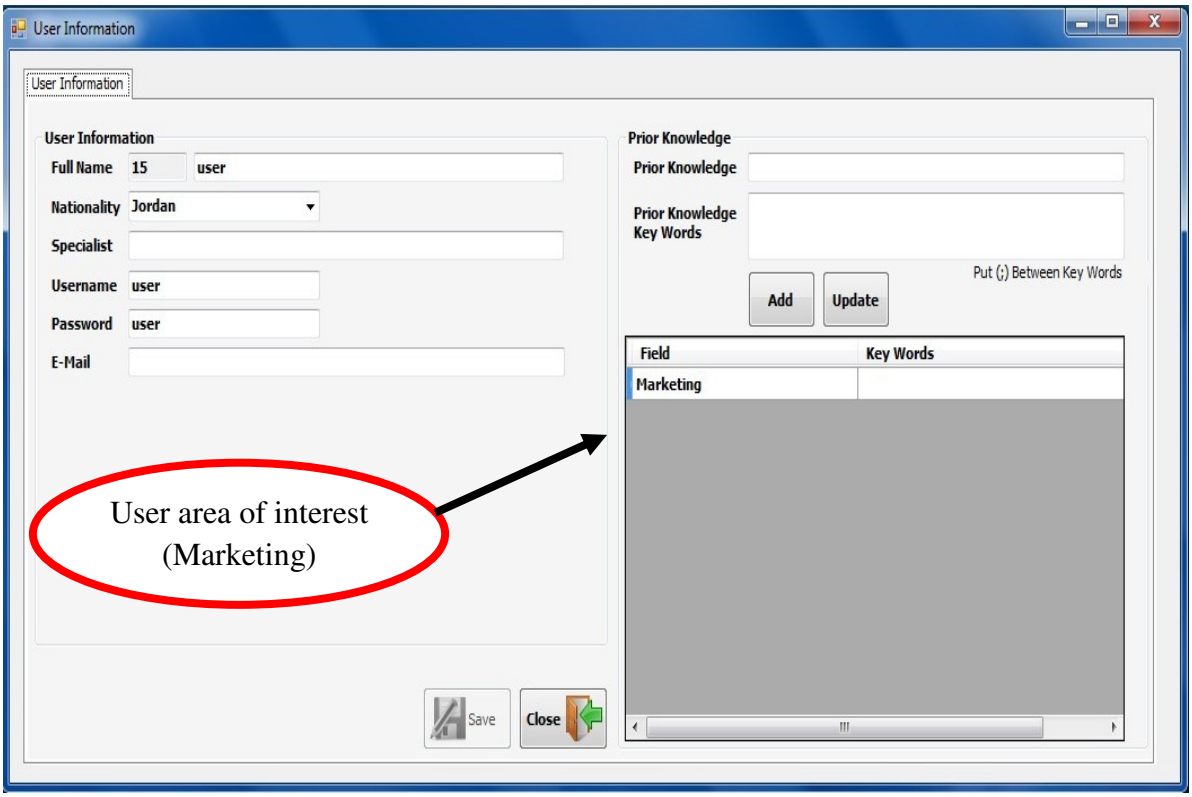


Figure (4.3) User Profile for Test1

From the previous test user area of interest is Marketing and the keyword entered is principles, so the user searching about any principles related with marketing domain. All links appear in web browsers such that GOOGLE or YAHOO related to user request, the first five links display below respectively.

a. Links by GOOGLE (first five from 60 related links)

[1] Marketing - SlideShare

http://www.slideshare.net/duongtit/marketing-1277880 Principles Of Marketing 1693 views Like. Marketing presentation 965 views Like. Pricing 4644 views Like. Pricing product-pricing-strategies 2601 views Like.

[2] Old Mutual | Interim Results 2012

http://financials.oldmutual.com/V2/results.jsp?result_id=12419&cat_id=12524 Market Consistent Embedded Value ... prepared in accordance with the recognition and measurement principles of International Financial Reporting Standards.

[3] Charlotte Samso - Danmark | LinkedIn

http://dk.linkedin.com/pub/charlotte-sams%C3%B8/0/986/590 Preparation of marketing materials for markets, incl.: principle design, key visuals, selling-in materials, etc. Tactical responsibilities for products for launch

[4] Excellent lab our market in Switzerland.-ST Galen Bodensee Area

http://www.sgba.ch/en/Business-Environment/Job-Market has an open, cross-border labor market and is attractive to workers from all over ... Switzerland's social security system is based the principles of solidarity.

[5] Impacts and Implementation of the Basel Accords - Scholarship ...

http://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1315&context=cmc_theses ,Apr 23, 2012 ... precondition for the core principles and Basel Accords. Market discipline refers to various aspects of financial markets such as the following: the

b. Links by YAHOO.COM

[1] Online TDM Encyclopedia - Market Principles

www.vtpi.org/tdm/tdm60.htm Cached Preface. A skilled dancer or athlete moves gracefully with minimal wasted effort: actions are anticipated, weight is shifted, muscles tension and relax, directing ...

[2] The Principles of the Free Market

www.free**market**monument.org Cached **Principles** of the Free **Market** (Draft Version) English 中文 EspaÃ±ol العربية Other Languages. 1) Individual Rights: "We are each created with equal ...

[3] What are the basic principles of a free market

wiki.answers.com/Q/...the_basic_**principles**_of_a_free_**market** Cached The Free **Market** Monument Foundation has done extensive research on what **principles** are most commonly associated with free **market** economics. Individual Rights. The ...

[4]Free market - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Free_**market** Cached:Economic systems|Concepts|Criticisms|See also The laissez-faire **principle** expresses a preference for an absence of non- **market** pressures on prices and wages, such as those from government taxes, ...

[5] Basic Stock Market Principles | eHow - eHow | How to Videos ...

www.ehow.com/info_7737357_basic-stock-**market-principles**.html Cached:

Every work day in the global marketplace, scores of financially savvy strangers meet to bid up or down the equity shares and bonds of publicly-listed companies ..

4.2 Test 2 : User enters one PK with one keyword Related to PK

In this level user has been entered PK and one keyword describe his/her area of interest, figure 4.4, 4.7 and 4.8 showing the result and UP for this level.

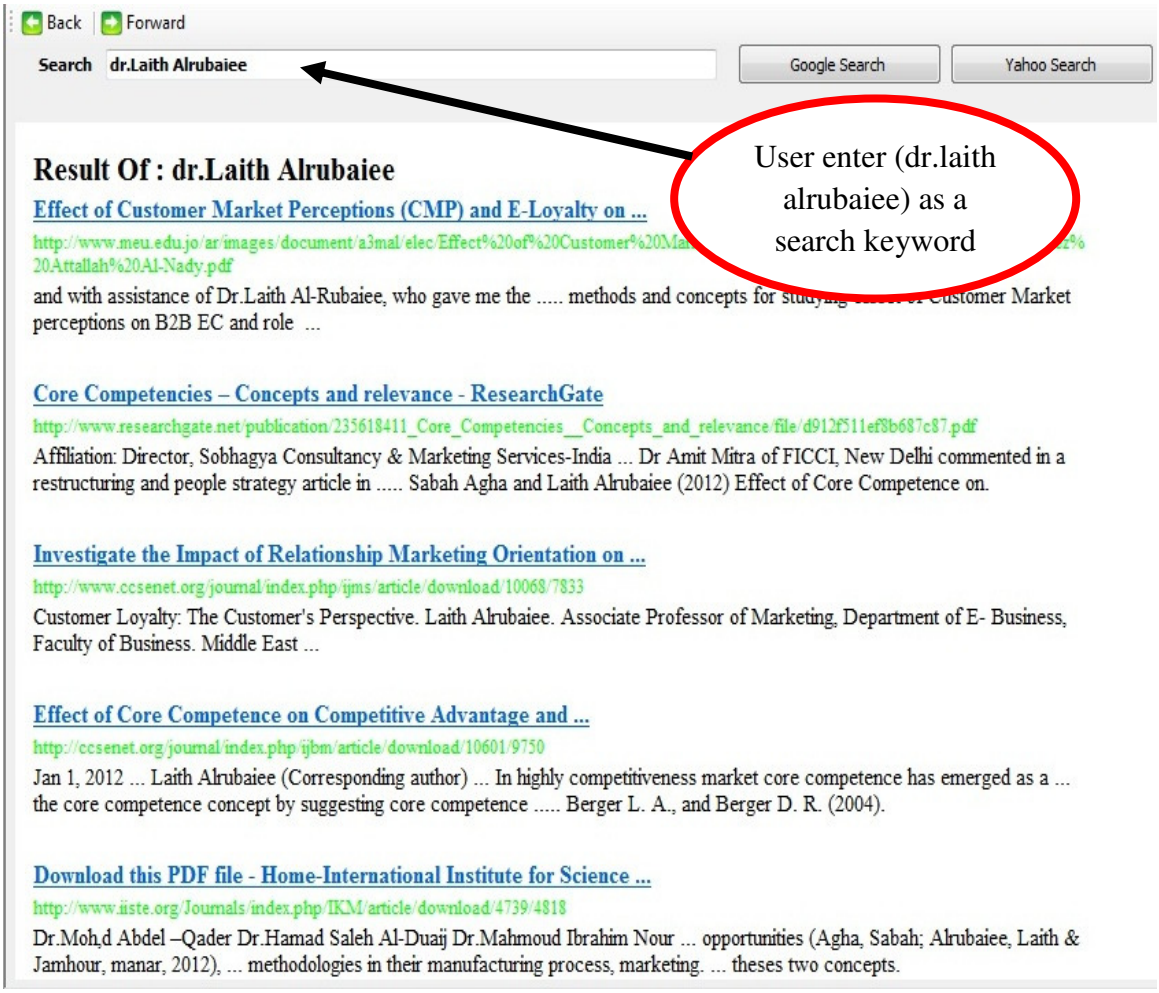


Figure (4.4) Results of Test2 Depend on GOOGLE API

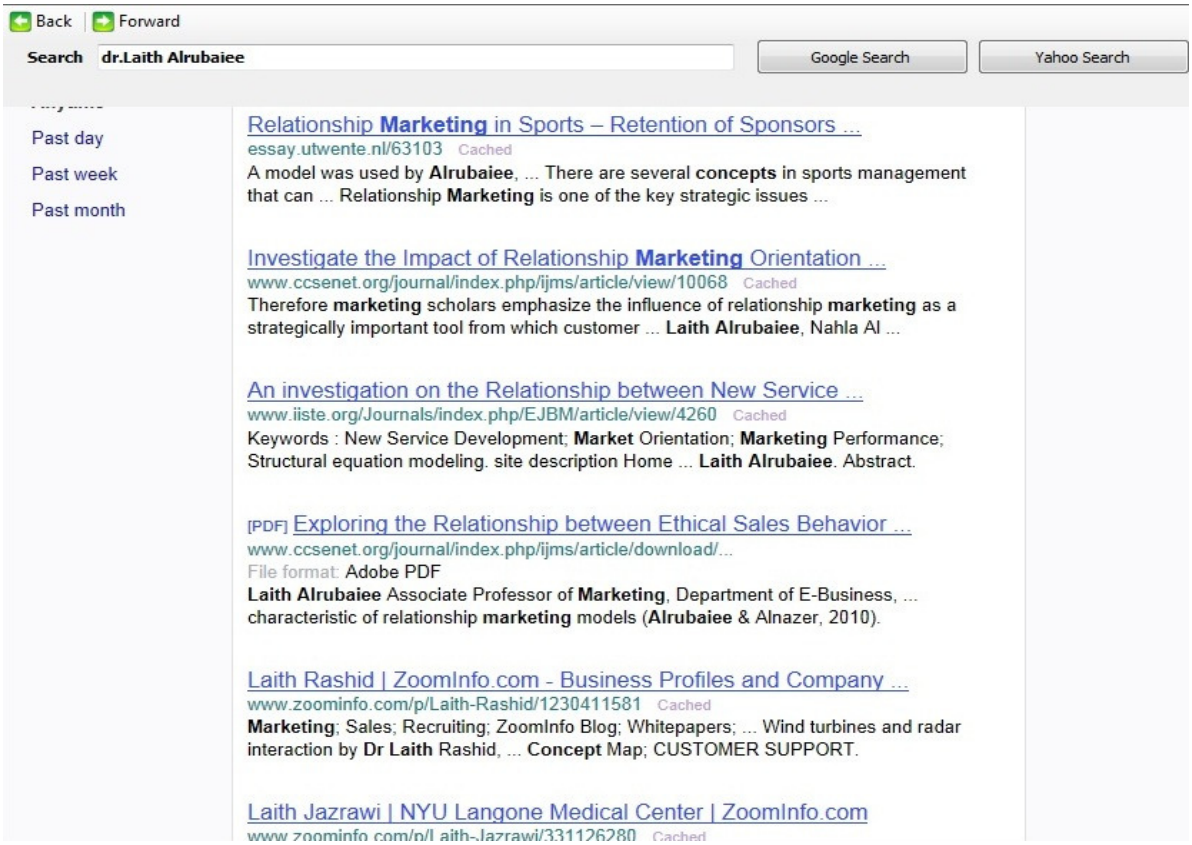


Figure (4.5) Results of Test2 Depend on YAHOO.COM

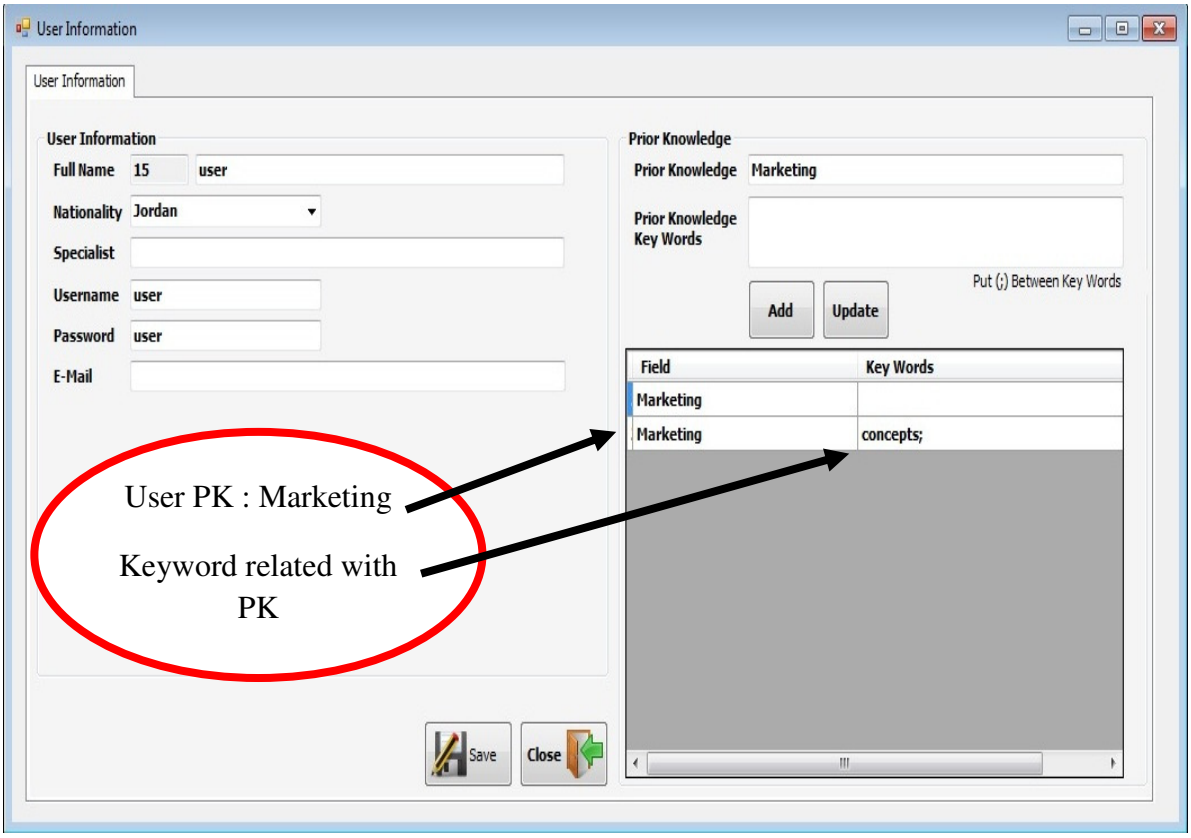


Figure (4.6) User Profile for Test2

From the previous test user PK is marketing and the specific topic which user need is concepts in marketing, each link appear depend on GOOGLE API or YAHOO.COM , talking about concepts in marketing.

4.3 Test 3 : User enters one PK with more than one keyword

In this level user have one PK but with more than one keyword, so the system will be depend on intersection rule, figure 4.7, 4.8 and 4.9 display the results and UP for this test level.

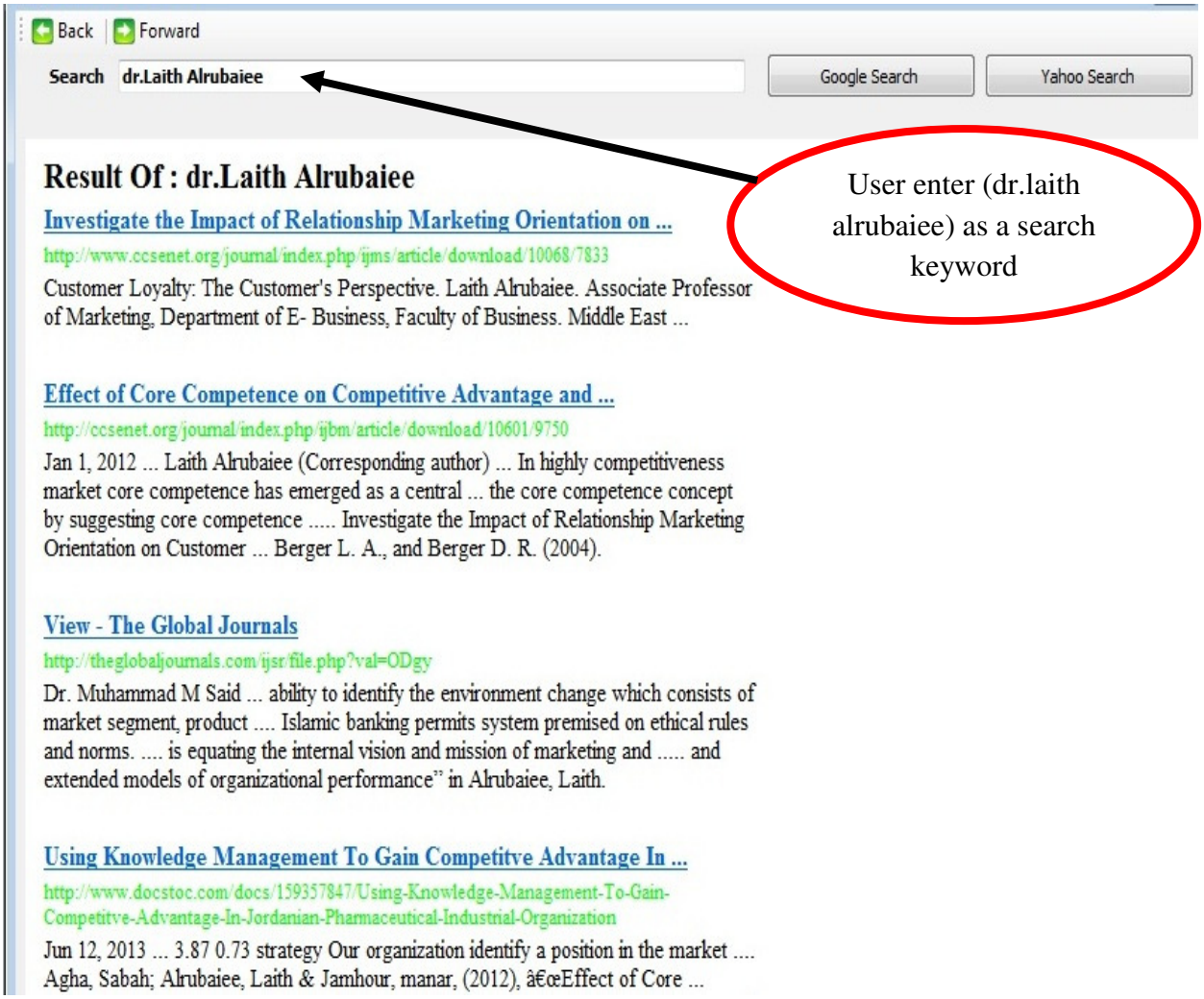


Figure (4.7) Results of Test 3 Depend on GOOGLE API

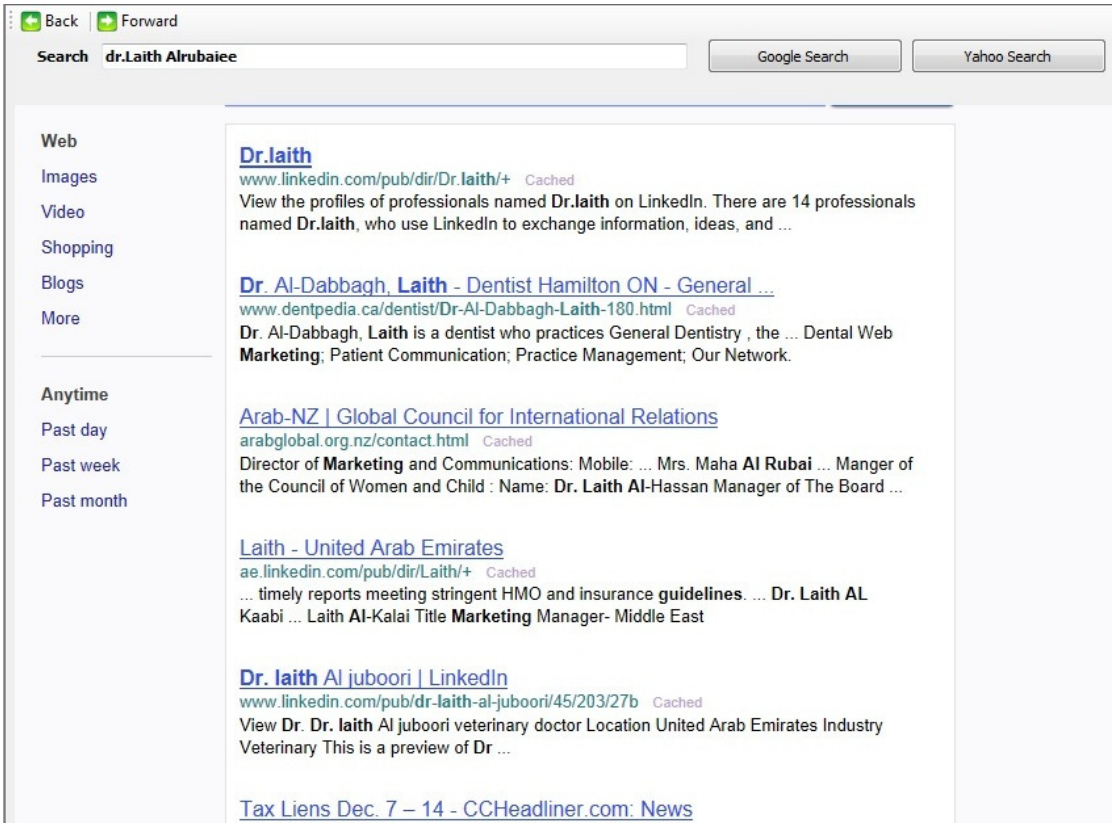


Figure (4.8) Results of Test 3 Depend on YAHOO.COM

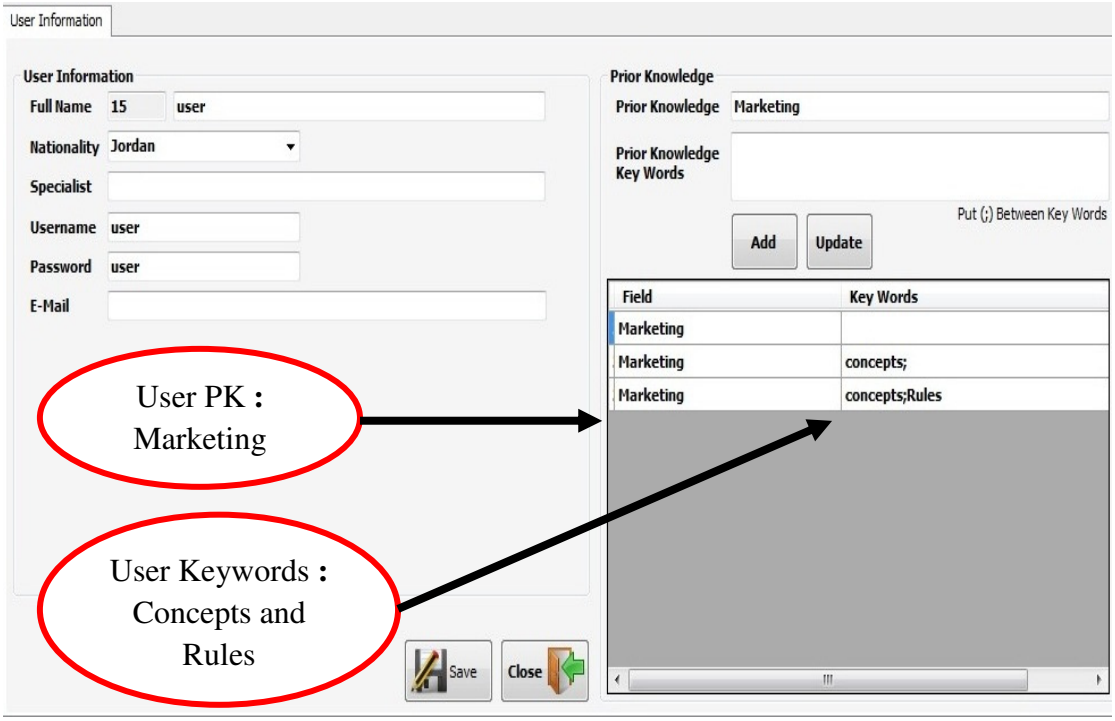


Figure (4.9) User Profile for Test 3

From the previous test we can see when the user entered ('Dr. Laith Alrubaiee') in the search text box in Test 2, the system retrieves different information when the user uses the same searching keyword in Test 3.

4.4 Test 4 : User enters two PK with more than keywords related with each PK

In this level user have many PK and many Keywords, figures 4.10, 4.11 and 4.12 show the results for this level.

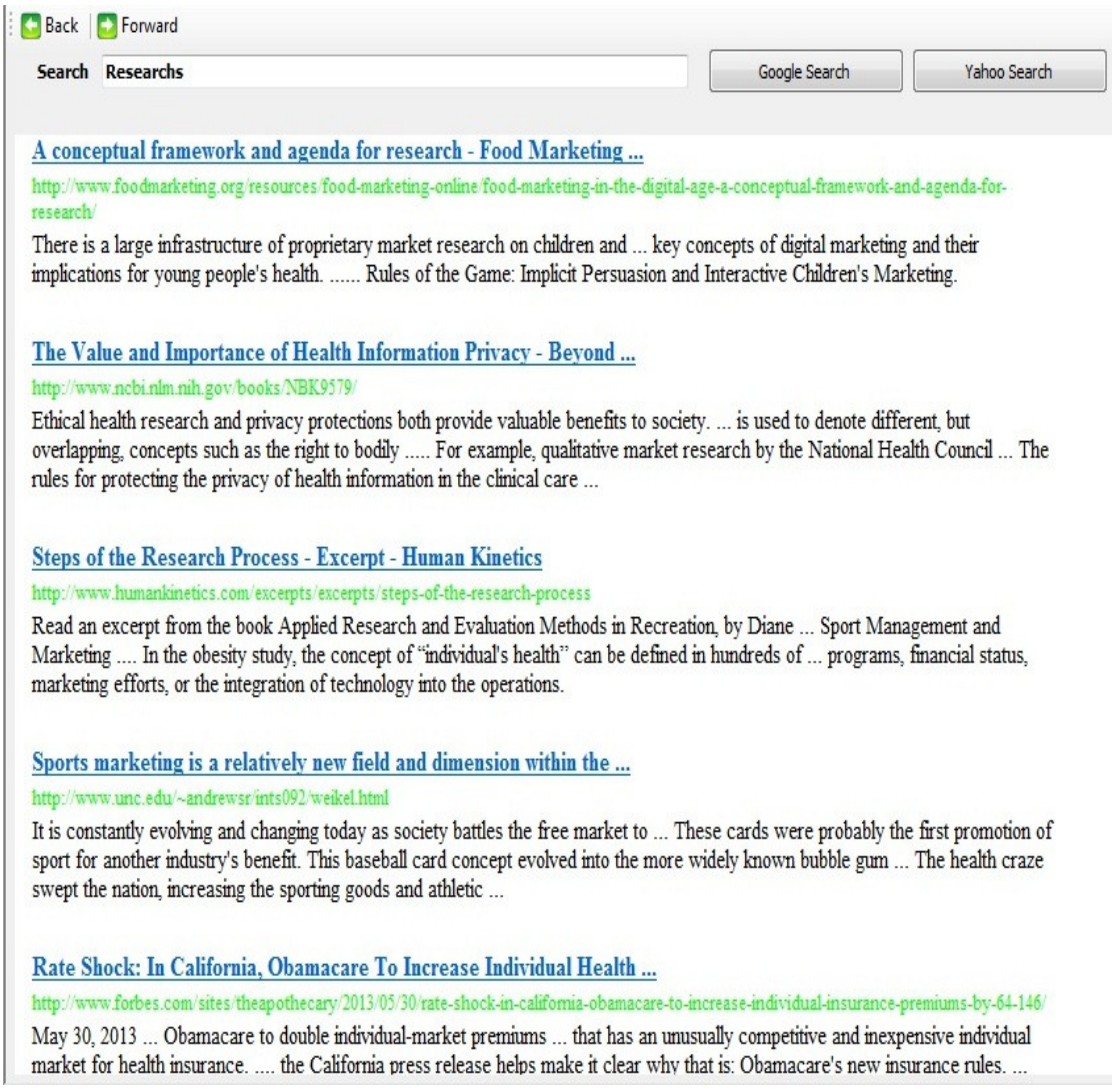


Figure (4.10) Result of Test 4 Depend on GOOGLE API

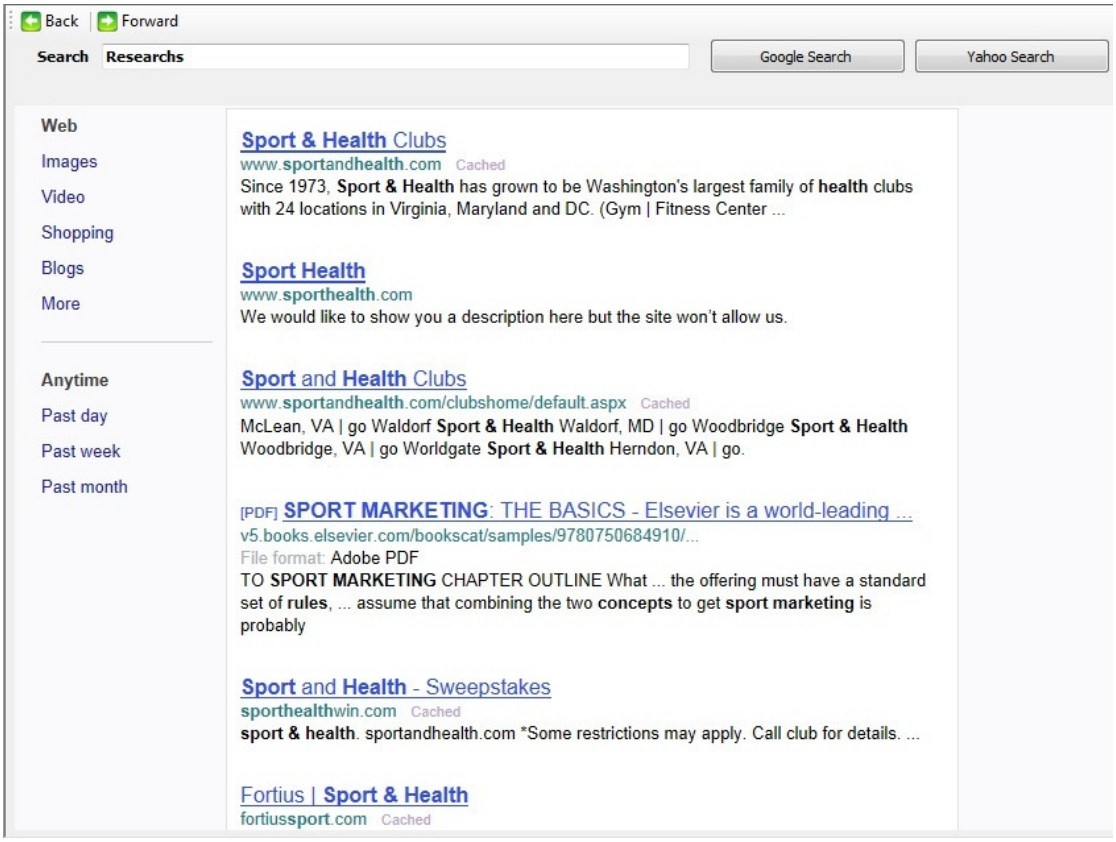


Figure (4.11) Results of Test 4 Depend on YAHOO.COM

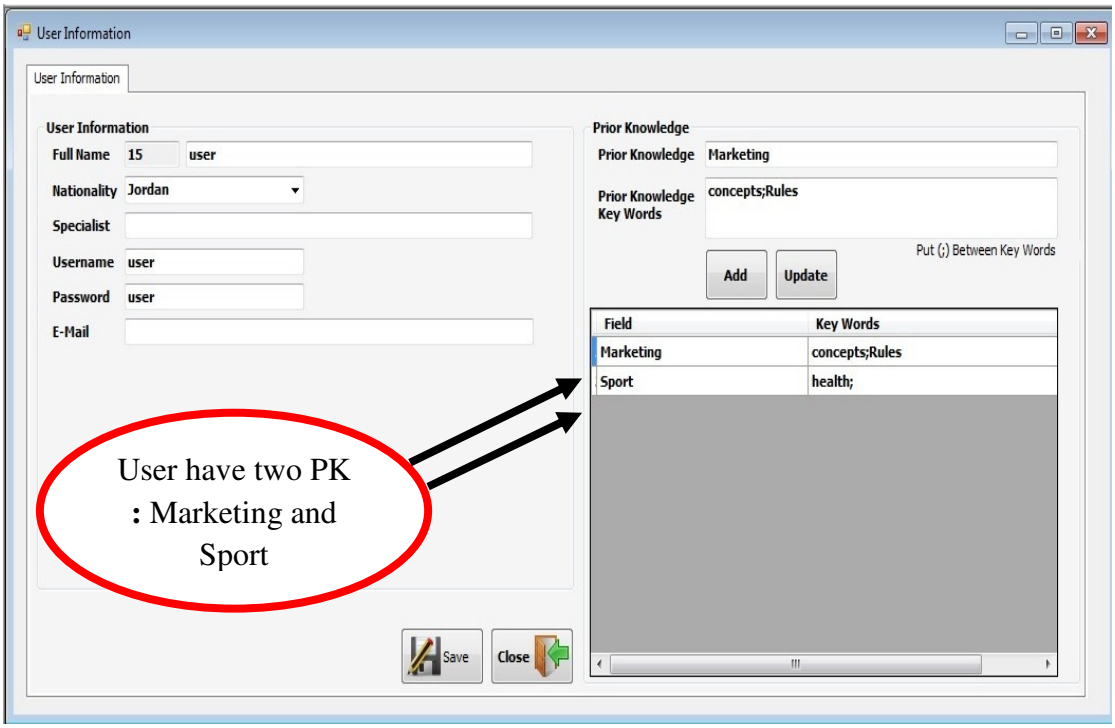
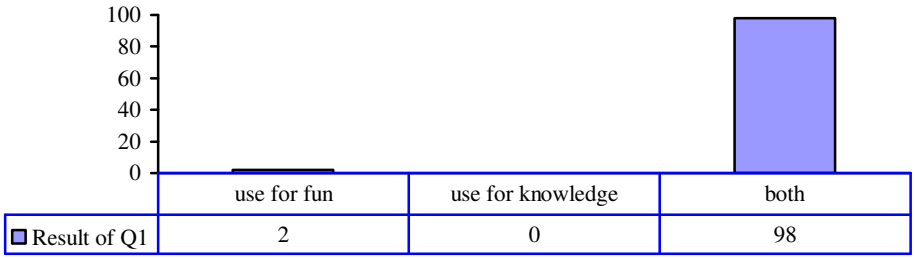


Figure (4.12) User profile for Test 4

4.5 Questioner Statistical Analysis

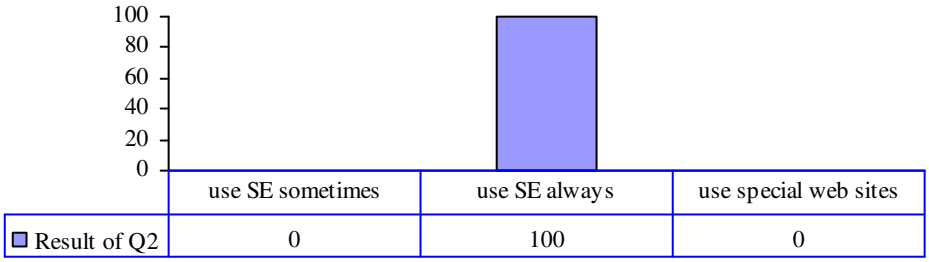
We review the questions and the results that we have obtained from the distribution of the questionnaire by using Bar Graph.

- Question No.1: Using Internet for fun, knowledge or both?



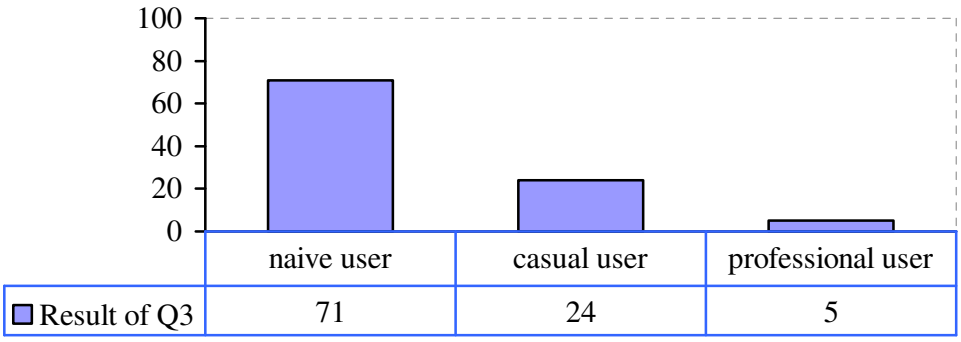
Bar Graph (4.1) Results of Question No.1

- Question No.2: Do you use Search Engines sometimes, always or use special web sites?



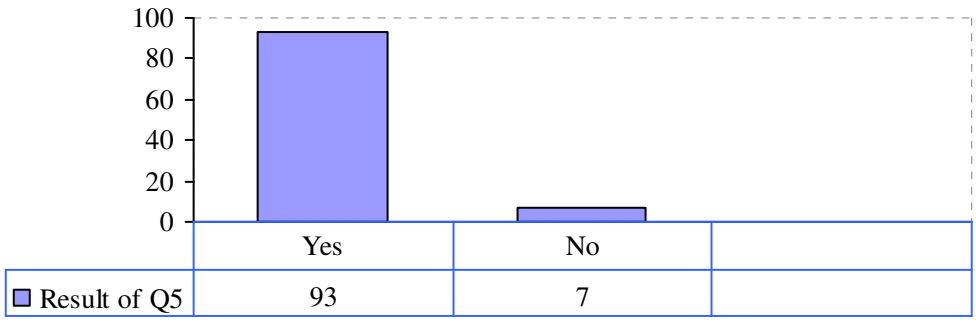
Bar Graph (4.2) Results of Question No.2

- Question No.3: The level of knowledge of search techniques?



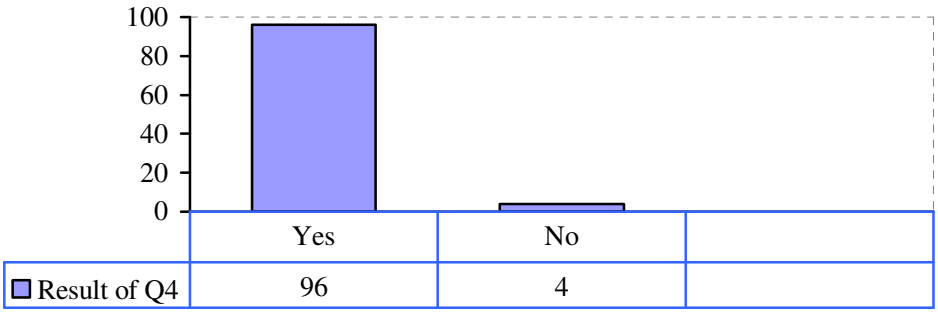
Bar Graph (4.3) Results of Question No.3

❖ Question No.4: Do you know the Boolean operators (+, OR , -) and how using it for retrieves useful information from WWW?



Bar Graph (4.4) Results of Question NO.4

❖ Question No.5 : Do you want to use an application that retrieving information from the WWW based just on areas of interest?



Bar Graph (4.5) Results of Question No.5

4.6 Summary

The proposed system doesn't represent any manipulation on Search Engines approaches for IR technology but represent the facility approach for naive WWW users to achieve their needed information with low effort. So the comparison below not for comparing this system with any Search Engine available now but for explaining the advantages of the proposed approach briefly in the next table.

Table (4.1) Advantages of using the proposed system

Properties	**Using The Proposed System**	**Using Search Engines Directly**
Number of Keywords Entered in Search Text Box	One keyword achieve the target	More than one keyword to achieve the target
Number of Links Appears	Few	Many
Speed of IR	Similar	Similar

From the previous testing and validation level (Test 4.1 to Test 4.2) we will conclude the following points :

1. Any user can be interesting with the proposed system facility after building their own profile.
2. The proposed rules learning user what the meaning of strong context and what is the importance of useful context for retrieving useful information related to user areas of interest.
3. The proposed system learning user how SE work.

Chapter Five

Conclusion and Future Works

5.1 Conclusion

In this thesis can conclude the following:

1. Information Retrieval Technology (IRT) based on two factors there are the Keyword and IR approach. Information Retrieval Systems (IRSs) used for retrieving online information (from WWW) and from static DB.
2. The proposed system is called Knowledge Based System (KBS) and uses one of knowledge representation approaches which is the Rule Base approach.
3. The best technique for connecting different user's areas of interest by using Boolean operators adding to the system as a set of rules.
4. Using Application Programming Interface (API) retrieves more related information.
5. The developed System, Knowledge-based system, can be used by the searching engines as facility available.

5.2 Future Works

There are many points for future works from the developed knowledge-based system to improve search technique and IR technology to collect better related results, such as:

1. Use API for all most popular search engines for collecting better results.

2. Build UP based on the most widely used approaches helps to reflect the user areas of interest.

3. Enhancing the user interface by adding more functions make search results ease to use directly, such that survey results in (PDF extension , DOC extension , PPT extension , ...etc.) without needing to write extension type by user, just selected from icon appear on the user interface.

References

A. Abdollahzadeh Barfourosh and M. L. Anderson, 2002, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, Institute for Advanced Computer Studies, and Department of Computer Science, University of Maryland, College Park, MD, USA.12-15.

Abdel-Badeeh M. Salem, Mohamed Roushdy and Rania A. HodHod, 2005, A case based expert system for supporting diagnosis of heart diseases, AIML Journal, Volume (5), Issue (1), March,33-35.

Addison Wesley, 2008, Search Engines : Information Retrieval in Practice, ISBN-13: 9780136072249, Addison-Wesley • Cloth, 552 pp Published 02/06/2009.

Ajith Abraham, 2005, Rule-based Expert Systems, Measuring System Design Handbook, John Wiley & Sons, Ltd. ISBN: 0-470-02143-8, Oklahoma State University, Stillwater, OK, USA.

Alexander Pretschner, 1998, Ontology Based Personalized Search, Master Thesis, Department of Electrical Engineering and Computer Science, University of Kansas , Germany.

Anglin, G., H. Vaez, and K. Cunningham, 2004, Visual representations and learning: the role of static and animated graphics, Ed. D. Jonassen. Hillsdale, NJ: Lawrence Erlbaum.

Athar Shaikh, Pravin Bhjantri, Shankar Pendse and V.K.Parvati, 2011, Profile Based Information Retrieval, International Journal on Computer Science and Engineering, Vol. 3 No. 4 April, Dharwad, India.

B. van Gils, H.A. Proper, P. van Bommel, and E. D, 2004, Schabell. Profile based retrieval on the world wide web. In P. de Bra, editor, Proceedings of the Conferential Informatiewetenschap (INFWET2003), pages 91.98, Eindhoven, The Netherlands, EU, 2003.

Bernard J. Jansen & Caroline M. Eastman, 2003, The Effects of Search Engines and Query Operators on Top Ranked Results, Information Technology: Coding and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference on 28-30 April, pages 135 - 139.

BO XING, 2006, Internet Search Engine Result Diversity, Relevance, Quality, and The Effects of Search Engine Optimization, PhD Thesis, Graduate Faculty of Texas Tech University, USA.

Clive Nettey, 2006, Link-Based Methods for Web Information Retrieval, MSc Thesis, at the Universities van Amsterdam.

Cornelius T. Leondes, 1998, Fuzzy Logic and Expert Systems Applications, 1st Edition,vol 6.

Dagobert Soergel, 1989, Important problems in information retrieval, Graduate Library School, University of Chicago.

David Hawking, 2006, Web Search Engines, Part1 : A behind - the - scenes look explores the data processing "miracle" that characterizes Web crawling and searching. CSIRO ICT Centre. pages 86-88.

Djoerd Hiemstra, 2009, Information Retrieval Models, Published in: Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, November. Netherlands.

Eileen Kowalski, 2004, Researching with Search Engine Operators, Pennsylvania State University, State of Pennsylvania, USA.

Eric Allen Engle, An Introduction to Artificial Intelligence and Legal Reasoning: Using(xTalk) to Model the Alien Tort Claims Act and Torture Victim Protection Act, published by Richmond Journal of Law & Technology, Volume XI, 4-7, Issue 1, Germany.

H.Araki, 2005, Information Retrieval Based on Text Classification, Federal University of Curitiba, Spain.

Hadeel Showket Al-Obaidy, (2009), Building Ontology Web Retrieval System Using Data Mining, Volume 4, 2-4, Computer Engineering Department College of Engineering Ahlia University, Bahrain.

Hang Li, 2011, Web IR-Through Introduction to Semantic Matching Project, MSRA WSM Lecture Series, Feb. 25, 11-19, Microsoft Research Asia.

Hsinchun Chen, 1995, Machine Learning for Information Retrieval: Neural Networks, Published by John Wiley & Sons. Inc., Journal of The American Society For Information Science. 46(3):194-216.

Hussein H. Owaied, 2012, New Overview Of Artificial Intelligence, Published in International Journal Of Academic Research. Vol. 4. No. 2. April.

J. Sarkar, (2001), Internet Search Tools and Techniques, Communication Support for Sustainable Development. Eds. Dipak De and Basavaprabhu Jirli, GangaKaveri Publishing House, Jangamawadi Math, Varanasi.

James Grimmelmann, (2007), The Structure of Search Engine Law, this Article is available for reuse under the Creative Commons Attribution 3.0 United States license, http://creativecommons.org/licenses/by/3.0/us/. All otherwise-undated web sites in footnotes were last visited on August 28.

James L. Crowley, 2010, Intelligent Systems: Reasoning and Recognition, Lecture 2, on Winter Semester 5 Feb, research center in Montbonno, France.

Kamlesh Kumar, Ashutosh Kumar Bhatt and Amarjeet Singh, 2012, Web-Based Support Systems For Information Retrieval, Journal of Information and Operations Management, ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1, 2012, pp-311-315 Available online at http://www.bioinfo.in/contents.php?id=55.

Kazunari Sugiyama, Kenji Hatano & Masatoshi Yoshikawa, (2004), Adaptive Web Search Based on User Profile Constructed without Any Effort from Users, Conference, May 17–22, New York, USA.

Lancaster and Warner 1993, Information Storage and Retrieval, Information Resources Press, US Mountain.

Lora K. Kaisler & Dennis O'Connor, 2006, Boolean Operators for Advance Search, Mathematics and science academy, EBook, 21st Information fluency.

Luciene Stamato Delazari, 2012, Semi-Automatic Semantic Data Classification Expert System to Produce Thematic, University of Parana, Geodetic Science Program, Curitiba, Parans, Brazil.

M. Sasikumar, S Ramani, S Muthu Raman, KSR Anjaneyulu & R Chandrasekar, 2007, A Practical Introduction to Rule Based Expert Systems, Narosa Publishing House, New Delhi.

Margaret Reid-Miller, 2010, Boolean Data Type, School of Computer Science Carnegie Mellon University. USA.

Mariana Maceiras Cabrera & Ernesto Ocampo Edye, 2010, Integration of Rule Based Expert Systems and Case Based Reasoning in an Acute Bacterial Meningitis Clinical Decision Support System, International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010, Montevideo, Uruguay.

Meredith Ringe Morris , Jaime Teevan & Katrina Panovich 2010, A Comparison of Information Seeking Using Search Engines and Social Networks, Microsoft Research, Redmond, WA, USA, Massachusetts Institute of Technology, Cambridge, MA, USA.

Michael Gordon & Praveen Pathak, 1998, Finding information on the World Wide Web: the retrieval effectiveness of search engines, Computer and Information Systems, University of Michigan, Ann Arbor, MI 48109-1234, USA.

Michael W. Berry, Susan T. Dumais and Gavin W. O'Brien, 1995,Using Linear Algebra for Intelligent Information Retrieval, Volume 37, Issue 4, 37(4) 573-595. (23 pages).University of Bristol, UK

Monika Henzinger, 2000, Link Analysis in Web Information Retrieval, Google Incorporated, Mountain View, California

Mylonas, D.Vallet, P. Castells, M. Fern´andez & Y. Avrithis 2004, Personalized information retrieval based on context and ontological knowledge, National Technical University of Athens, United Kingdom.

Myriam Hadjouni, Mohamed Ramzi Haddad, Hajer Baazaoui, Marie Aude Aufaure, & Henda Ben Ghezala, 2009, Personalized Information Retrieval Approach, Laboratory RIADI-GDL, National School of Computer Sciences, University of Manouba, la Manouba, Tunisia.

Natasha Alechina, 2012, Knowledge representation and reasoning, Lecture 1: Introduction, School of Computer Science University of Nottingham, NG8 1BB UK.

Priti Srinivas Sajja & Rajendra Akerkar, 2010, Text Book : Knowledge-Based Systems : Knowledge Based System Components, India.

Robert M. Losee & Manning Hall 1998, Comparing Boolean and Probabilistic Information Retrieval Systems Across Queries and Disciplines, Published in J. of the American Society for Information Science 48 (2) 1997, 143–156. June 28,, USA Chapel Hill.

S. Quaglini, P. Barahona & S. Andressen , (2001), Artificial Intelligence in Medicine, Proceedings of 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001, Cascais, Portugal, Springer, July, 2001.

Saad Hussein Fathalla, 1999, Principles of statistics and statistical methods, Mafraq-Jordan.

Susan Gauch , Mirco Speretta , Aravind Chandramouli , and Alessandro Micarelli, 2007,User Profiles for Personalized Information Access, Adaptive Web, LNCS 4321, pp. 54 – 89, 2007. Springer-Verlag Berlin Heidelberg.

Srinath Srinivasa & P C P Bhatt, 2002, Introduction to Web Information Retrieval: A User Perspective, Indian Institute of Information Technology, Bangalore.

Terre Haute, 1999, Keyword And Boolean Searching ,Library Instruction, Indiana State University Libraries.

Tim Finin, James Mayfield, Anupam Joshi, R. Scott Cost & Clay Fink, 2005, Information Retrieval and the Semantic Web, Baltimore County Baltimore MD 21250, USA.

Venkat N. Gudivada, Vijay V. Raghavan, William I. Grosky & Rajesh Kasanagottu, 1997, Information Retrieval On The World Wide Web, University of Missouri, USA.

William. Wood, Ann. Houston, Robert.Kuhns, Paul. Martin &n Stephen. Green, 2000, Linguistic Knowledge can Improve Information Retrieval, Page 45-94.

William Siler & James J. Buckley, 2005, Fuzzy Expert Systems And Fuzzy Reasoning, Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published simultaneously in Canada. Includes bibliographical references and index. ISBN 0-471-38859-9

Yogendra Kumar Jain & Sandeep Wadekar, 2011, Classification-based Retrieval Methods to Enhance Methods to Enhance Information Discovery on the Web, International Journal of Managing Information Technology,3(1), 33 - 44.

Zhongming Ma, Gautam Pant & Olivia R. Liu Sheng, 2007, Interest-Based Personalized Search, The University of Utah. ACM Transactions on Information Systems, Vol. 25, No. 1, Article 5, Publication date: February.

Appendix

- Appendix (A)
  a. (English Questioner Form)

Middle East University

Faculty of Information Technology

Department of Computer Science

Mr., Mrs.

The researcher conducted a study entitled " **Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge**" and as one of the requirements for obtaining a master's degree in computer science from the Middle East University, and under the supervisor Dr. Hussein H. Owaied, to accomplish this task, please fill out this questionnaire accurately and objectively note that the information will be confidential for the purposes of scientific research.

Yours sincerely...

Researcher

Ghaith Alkubaisi

Middle East University

- **Part 1**

❖ Personal Data

Faculty:

Course :

Areas of Interest:

Age:

- **Part 2**

Please put (X) mark front the appropriate answer

❖ Question 1 : Using Internet for fun, knowledge or both

The Phrase	Answer
Using for fun	
Using for knowledge	
Using for fun and Knowledge	

❖ Question 2 : Do you use Search Engines

The Phrase	Answer
Using search engines sometimes	
Using search engines always	
Using special web sites	

❖ Question 3 : The level of knowledge of search techniques

The Phrase	Answer
Naive User	
Casual User	
Professional User	

❖ Question 4 : Do you know the Boolean operators (+, OR , -) and how using it for retrieves useful information from WWW

The Phrase	Answer
Level of knowledge nil	
Level of knowledge weak	
Level of knowledge average	
Level of knowledge advance	

❖ Question 5 : Do you want to use an application that retrieving information from the WWW based just on areas of interest

The Phrase	Answer
Want	
Want strongly	
Don't want	
Don't want strongly	

- Thank you, please indicate any aspirations or appropriate any comments you wish on the subject.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

b. (Arabic Questioner Form)

جامعة الشرق الاوسط

كلية تكنولوجيا المعلومات

قسم الحاسوب

السيد الفاضل , السيدة الفاضلة

تحية طيبة وبعد...

يقوم الباحث بأجراء دراسة بعنوان " تصميم وتنفيذ نظام مبني على المعرفة لاسترجاع النص بالأعتماد على سياق النص والمعرفة المسبقة للمستخدم" , وذلك كأحد متطلبات الحصول على درجة الماجستير في علم الحاسوب من جامعة الشرق الاوسط, وتحت اشراف الدكتور حسين هادي عويد.

ولإنجاز هذه المهمة, يرجى التكرم بتعبئة هذه الاستبانة بدقة وموضوعية علماً بأن المعلومات ستقدم بسرية ولأغراض البحث العلمي.

وتفضلوا بقبول فائق الاحترام,,,

الباحث

غيث عبدالستار الكبيسي

جامعة الشرق الاوسط

- **القسم الاول**

❖ البيانات الشخصية

الاختصاص :

المرحلة :

مجالات الاهتمام :

العمر :

- **القسم الثاني**

يرجى وضع علامة (X) امام الاجابة المناسبة

❖ السؤال الاول : مجال استخدامك للأنترنت

العبارة	الاجابة
تستخدم الانترنت للمتعة	
تستخدم الانترنت للحصول على معرفة	
تستخدم الانترنت للمتعة وللحصول على معرفة	

❖ السؤال الثاني : هل تستخدم محركات البحث عبر الانترنت

العبارة	الاجابة
تستخدم محركات البحث احيانا	
تستخدم محركات البحث دائما	
تستخدم مواقع خاصة	

❖ السؤال الثالث : مستوى معرفتك بتقنيات البحث واسترجاع المعلومات عبر الانترنت

العبارة	الاجابة
مستوى مبتدأ	
مستوى متقدم	
مستوى محترف	

❖ السؤال الرابع : هل تعرف العمليات المنطقية (AND,NOT,OR) المستخدمة لاسترجاع المعلومات المناسبة من الانترنت

العبارة	الاجابة
مستوى المعرفة معدوم	
مستوى المعرفة ضعيف	
مستوى المعرفة متسوسط	
مستوى المعرفة متقدم	

❖ السؤال الخامس : هل ترغب بأستخدام تطبيق يعمل على استرجاع المعلومات من الانترنت ضمن مجالات اهتمامك فقط

العبارة	الاجابة
ارغب بشدة	
ارغب	
لا ارغب	
لا ارغب بشدة	

- شكرا لكم , الرجاء بيان اي تطلعات او اي ملاحظات مناسبة ترغبون بها حول الموضوع .

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

❖ Appendix (B)

- **اسماء محكمي الاستبانة**

د. حسين عويد الشمري	**جامعة الشرق الاوسط**
ا.د علاء حسين الحمامي	**جامعة عمان العربية**