ARAB ACADEMY FOR SCIENCE, TECHNOLOGY AND MARITIME TRANSPORT (AASTMT)

College of Computing and Information Technology

Department of Computer Science

ENHANCING WEB INFORMATION RETRIEVAL USING IMPROVED RANKING AND A PARALLEL CORPUS TECHNIQUE

By

ALI M. NABIL ALLAM

A thesis submitted to AASTMT in partial
fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE

IN

Computer Science

Supervisors

Dr. Mohammed Sakre
High Institute of Computers &
Information Technology
Al-Shorouk Academy

Prof. Dr. Mohamed Kouta
Head of MIS Department
Arab Academy for Science,
Technology & Maritime Transport

May 2009

APPROVAL OF DEFENSE COMMITTEE

We certify that we have read the present work and that, in our opinion, it is fully adequate in scope and quality as a thesis towards the partial fulfillment of the Master Degree requirements in

Specialization: Computer Science

From

College of Computing and Information Technology (AASTMT)

Date: 24-May-2009

"Enhancing Web Information Retrieval using Improved Ranking and a Parallel Corpus Technique"

By

Ali Mohamed Nabil Allam

Supervisors:

Name: Prof. Dr. Mohamed Mahmoud Kouta.
Position: Head of Management Information Systems Department, AASTMT.
Signature:

Name: Dr. Mohammed Mahmoud Sakre.
Position: Lecturer at the High Institute of Computers and Information Technology, Al-Shorouk Academy
Signature:

Examiners:

Name: Prof. Dr. Ahmed Mohamed Hamad.
Position: Dean of Faculty of Informatics and Computer Science, BUE.
Signature:

Name: Prof. Dr. Zain Al-Din Mohamed Abdel Hady.
Position: Head of the Department of Libraries and Information, Helwan University
Signature:

PUBLISHED WORK

[1] Sakre, M., Kouta, M. & Allam, A., 2009. Weighting Query Terms using WordNet Ontology. *International Journal of Computer Science and Network Security (IJCSNS)*. Korea, Seoul. 9 (4), p.349-358. Accessible at: http://paper.ijcsns.org/07_book/html/200904/200904047.html

[2] Sakre, M., Kouta, M. & Allam, A., 2009. Automated Construction of Arabic-English Parallel Corpus. *International Journal of Computer Science and Network Security (IJCSNS)*. Korea, Seoul. (Accepted for publication).

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to all those who gave me the support to complete this research. Foremost, I wish to express my deepest gratitude and recognition to my supervisor, Prof. Dr. Mohamed Kouta. His encouragement, advice and example, on both a professional and personal level, provided a sterling impetus on my own professional development for which I will always be indebted.

I also owe Dr. Mohammed Sakre great thanks and appreciation for his continuous support and valuable suggestions in this thesis and in the published papers.

My special thanks go to Dr. M. Asaad Elnidani, Dean of College of Management and Technology at the Arab Academy for Science, Technology and Maritime Transport (AASTMT). He deserves gratitude for creating a unique atmosphere of research, work, trust and responsibility.

Special appreciation is also due to Dr. Mohamed Aborizka, Head of E-commerce Department at AASTMT, for providing me with additional encouragement in my research work through the high standards and perseverance he set in his own scientific endeavors.

I also like to extend my gratitude to both Dr. Eng. Kadreya Abou-Sayed and Dr. Eng. Ahmed Abou-Sayed, CEO of Advantek Corporation (Houston, Texas), for their efforts in facilitating the procedures of obtaining the license of WT2g collection from the University of Glasgow (UK). Their incredible efforts have made the completion of this research possible.

Finally, saving the best for last, I wish to give my heartfelt gratitude to my father, may Allah bless his soul, and my mother for their encouragement and support throughout the course of my graduate studies. Their continuous and endless support will be deeply cherished all my life through.

Thanks to all.

Ali Allam

ABSTRACT

Information domains, such as the web, have enormous information content. Therefore, the task of extracting information relevant to a particular topic or trying to predict what sort of information a user is seeking is not a trivial task; thus finding information relevant to a particular area of interest can be sometimes inconvenient and frustrating.

However, the goal of any information retrieval system is to retrieve accurate results in response to a query submitted by the user, and to rank these results according to their relevancy, with the ability to cross all language barriers. In order to achieve this goal, the proposed system architecture presented the following main components: concept-based term weighting (CBW), context-based matching (CM) and automated parallel corpus construction.

Concept-based term weighting (CBW) employed the conceptual information found in the WordNet ontology to determine the significance of query terms without depending on document collection statistics.

Context-based Matching (CM) showed how document term significance could be derived by interpreting context in queries and documents. Unlike Term Frequency (TF), which requires a term to occur frequently within a document to be significant, CM considered a term significant even if it did not occur frequently within a document.

The system also proposed a technique that constructed an Arabic-English parallel corpus automatically, through web mining. The technique succeeded to construct the parallel corpus through mining an Egyptian news website.

Finally, the system was tested using the *WT2g* web document collection as a benchmark under the *Terrier* package. Terrier was used first to index the WT2g collection, and then to retrieve the relevant documents in response to the topics of TREC, and finally to evaluate and compare the presented techniques against the traditional ones.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION.

This chapter provides a background overview of the unique nature of information on the web. It also covers the challenges that motivate web information retrieval research, as well as the objectives of the research and how they can be accomplished. Finally, it describes the organization of the remaining chapters of the thesis.

1.1. BACKGROUND OVERVIEW:

The World Wide Web is a popular and interactive medium to spread information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to those situations, web users are currently drowning in information and facing extraction and retrieval challenges (Kosala & Blockeel 2000).

The task of extracting information relevant to a particular topic, or trying to predict what sort of information a user is seeking is not trivial task, and therefore finding information relevant to a particular area of interest can be sometimes inconvenient and frustrating as well. People either browse or use the search service when they want to find specific information on the web. When a user uses a search service he or she usually inputs a simple keyword query and the query response is a list of pages ranked based on their similarity to the query (Namjoshi 2004).

Traditional information retrieval techniques rely on measures such as the frequency of a word in a given document, or the hyperlink connectivity of that particular web document. This approach may not necessarily bring out the important words or terms in a document and thus could be less effective while returning search results for queries.

1.2. CHALLENGES AND MOTIVATIONS:

The ultimate challenge for web information retrieval is to provide improved systems that retrieve the most relevant information available on the web to satisfy a user's information need. Quite clearly, the motivation to provide improved information retrieval systems is abundant. Specifically, the operative challenges motivating researchers in web information retrieval

include problems relating to the data and the user. Hereafter are some of the challenges of extracting information over the Web:

i. Amount of information:

It is fair to assume that the web contains information about almost any topic known to us. According to a post of the Official Google Blog, the first Google index in 1998 had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, Google systems that process links on the web found 1 trillion unique web pages on the web at once, of which 40 billion web pages only were indexed (Alpert & Hajaj 2008). Because the web consists of this enormous number of web pages, an answer to a user's query may consist of thousands, millions, or even billions of potentially relevant documents. Addressing the challenge of how the ranking of documents is performed is of great significance to the retrieval process.

ii. Multilingual contents:

The globally interconnected information infrastructure is known as the World Wide Web. However, for someone who reads only English, it is English-Wide-Web. A reader of only Arabic reads only the Arabic-Wide-Web. The challenge of crossing the language barriers will be discussed in section (2.8).

iii. Heterogeneity:

Web documents are typically heterogeneous HTML pages containing textual and multimedia contents (e.g. audio, video, images, or maps). Moreover, there are no standard rules for creating web documents, resulting in the wide variety of web documents.

iv. Life span of information on the web:

The freedom for anyone to publish information on the web at anytime and anywhere means that information on the web is highly dynamic. Some websites, such as news websites, have extremely volatile content, while others have information that keeps appending, such as educational websites or regular corporate websites.

v. Hyperlink connectivity:

The web is a collection of hyperlinked web documents that contain pointers to each other, creating communities of distributed interlinked information. This challenge will be discussed in more details in section (2.4.6).

vi. Users' behavior:

According to an analytical study carried out in 2000, users usually submit weak and short queries of 2.21 terms in average to express their information need (Jansen et al 2000). In addition, this study has also shown that when browsing the search results, 75% of the users get bored and do not usually browse beyond the second page of results, which raises the importance of documents ranking.

1.3. AIMS AND OBJECTIVES OF RESEARCH:

The objective of this research is to present web information retrieval techniques that would improve retrieval and ranking effectiveness in response to a submitted query with the ability to bridge the language gap between the user and the web. Therefore, the research questions addressed are:

(i) How could the submitted query terms be weighted using the conceptual information provided through ontologies?

(ii) How could the terms occurring in the context of the documents improve retrieval?

(iii) How could the term and document weighting measures be used to improve the ranking of documents?

(iv) How could the web provide parallel bilingual texts that can be used in crossing the language barriers between the user and the web?

Therefore, the main aims of this research, which address the abovementioned questions respectively, are to:

(i) Investigate a conceptual weighting technique for weighting query terms using ontologies to determine the significance of query terms.

(ii) Investigate a contextual technique for weighting document terms by interpreting the context in which a term appears in a document.

(iii) Investigate a technique for fusing the query and document weighting measures to calculate the rank score of a document.

(iv) Investigate and present a technique for mining the web to construct bilingual parallel texts.

1.4. RESEARCH METHODOLOGY:

All the techniques presented in this research attempt to enhance web information retrieval systems by using improved term weighting and document ranking techniques. In addition, the research presents a web mining technique to construct bilingual parallel texts that serve as good source of statistical translation models. Therefore, the methodology followed in this research addresses the following main aspects:

- Weighting query terms using conceptual information found in ontologies.
- Weighting document terms based on the context that they appear within.
- Ranking the retrieved documents according to their relevancy.
- Constructing a bilingual parallel corpus.

The aforementioned aspects are accomplished using a number of proposed techniques that are very different from traditional ones. Therefore, each traditional technique is substituted by its corresponding proposed technique, and consequently the proposed technique is evaluated and analyzed in order to provide a comparison that concludes the points of strength and weakness.

However, the research methodology employs one of the TREC web tracks, namely WT2g, which acts as a benchmark to provide robust evaluations and consistent comparisons of the presented techniques against the traditional ones.

1.5. ORGANIZATION OF THE THESIS:

This thesis is organized into five chapters followed by references and appendices. Chapter (2) is mainly concerned with reviewing pertinent literature related to various aspects of web information retrieval research. It reviews document classification techniques, term significance measures, and document ranking algorithms. The literature review also presents retrieval accuracy measures, concept-based retrieval, information personalization and cross-language approaches used for information retrieval.

Chapter (3) presents the system architecture and describes each of its components, as well as the techniques used to combine these components. These components are: concept-based term weighting, context-based matching, weighted document ranking, and parallel corpus construction.

Chapter (4) lays out the experimental results obtained during the course of this work, while chapter (5) forms the conclusion obtained by the research and outlines a direction for the work that could be done in the future.

A section listing the references that are cited in the thesis follows chapter (5), along with the bibliography used in preparing this research but not cited in the thesis. The appendices follow both, the references and bibliography, and presents additional information that is not directly included in the thesis.

CHAPTER 2

LITERATURE REVIEW.

This chapter gives a review of the pertinent literature related to various aspects of web information retrieval research in the past ten years. It discusses the related work in this field that run along a parallel vein to this research. It reviews document classification techniques, term significance measures, and document ranking algorithms. The literature review also presents retrieval accuracy measures, concept-based retrieval, information personalization and cross-language approaches used for information retrieval.

2.1. WEB INFORMATION RETRIEVAL SYSTEM:

Information retrieval systems appear in the web with the purpose of managing, retrieving and filtering the information available on the web. The two main technologies used for web information retrieval are the web directories and the search engines.

Web directories are the ontology of the web, where the most relevant documents are classified according to topic, placing quality above quantity of documents. On the other hand, search engines index, ideally, the whole documents available in the Web, placing quantity above quality of its contents. However, because the information on the web is huge, volatile and distributed, search engines will have to index an incredibly high amount of information and to update the frequent changes on the documents.

Figure (2.1) shows the most important components of a typical web information retrieval system. Web crawlers, which are also sometimes called spiders or robots, are programs responsible for collecting the web pages and storing them in the local document repository. Such a repository may then serve the needs of a web search engine. Crawlers start with a comprehensive set of root URLs called seed pages, and then follow the links on these pages recursively to find additional pages and append them to the repository without duplication. The indexer then processes those novel documents and typically creates an inverted term index and a document index. Once the indices are built, document ranking can be performed to weight terms by using the information stored in the indices; and as a result, a set of relevant ranked documents is returned to the user.

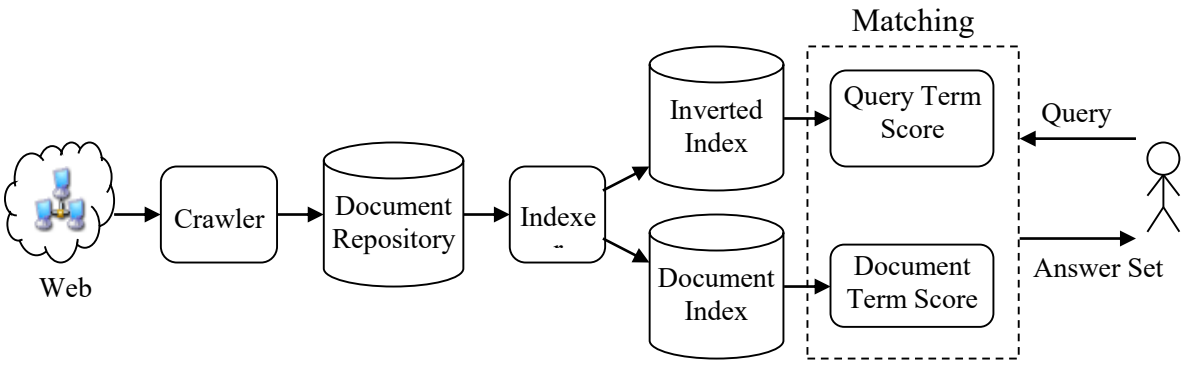

Figure 2.1: Architecture of a web information retrieval system

2.2. CLASSIFICATION TECHNIQUES:

The term classification is often used interchangeably with the terms cataloguing, categorizing, taxonomy and ontology. Classification structures are useful for organizing and finding information. There are various classification techniques, which will be discussed in the following sections. The right use of classification structures in the web information architecture can serve as an effective tool for information retrieval (Bates 2002).

2.2.1. Hierarchical Classification

Hierarchies are best when the entities in question are viewed in such a way that they have one dimension of classification, in which hierarchies divide and re-divide things into groups. Figure (2.2) is an example of a hierarchical classification which shows that its key feature is inheritance.

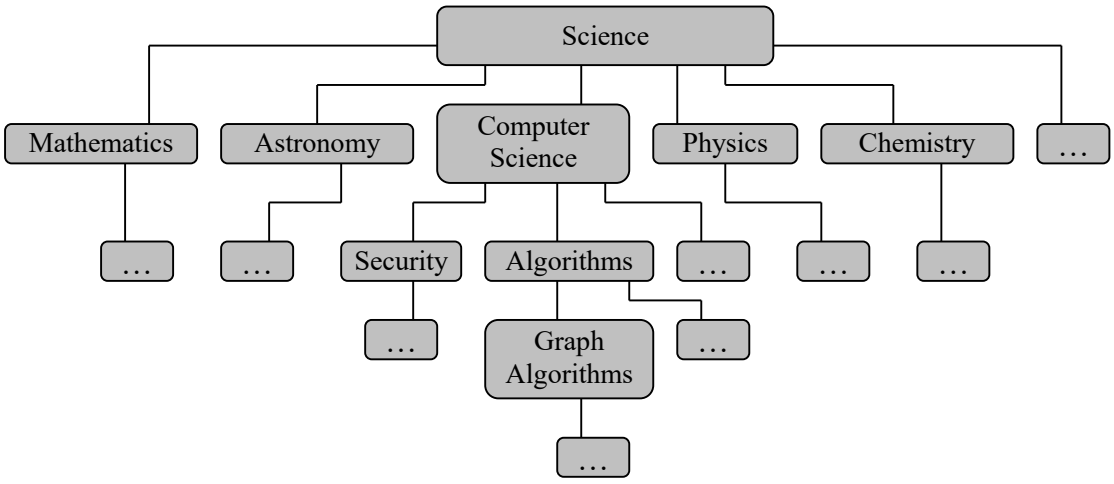

Figure 2.2: Example of hierarchical classification

The most three widely used universal hierarchical classification schemes are Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC), and the classification scheme devised by the Library of Congress (LLC). For example, DDC is used in BUBL Information Service Browsing (http://bubl.ac.uk), UDC is used in the browsing section of SOSIG (www.sosig.ac.uk), and LLC is used in CyberStacks (www.public.iastate.edu). In DDC and UDC, the whole universe of knowledge is firstly divided into ten major classes, then each class is divided again into ten divisions forming a hundred divisions which are again divided into thousands subdivisions. Thus the knowledge structure of DDC and UDC almost cover the whole field of knowledge. In contrast, LLC is built on 21 major classes, each class being given an arbitrary capital letter between A and Z, with 5 exceptions which are (I,O,W,X, and Y). LLC notations are composed of letters and numbers, where the capital letters are used for main class and subclass notations while numerals are used for subdivisions further down the hierarchies. For example, HB1-3840 is the LLC notation for "Economic Theory".

2.2.2. Tree Classification

Trees divide and subdivide its classes based on specific rules for distinction just like hierarchies, but do not assume the rules of inheritance. In a tree, entities have systematic relationships but not the generic 'is-a kind of' relationship. Examples: Trees for the chain of command in the army: Generals, Brigadiers, Colonels, Majors, Lieutenants, Sergeants; other examples of trees are such as geographic areas, organs of a body and parts of a vehicle.

2.2.3. Paradigm Classification

In contrast to hierarchies and trees, a paradigm (or matrix) is a two-dimensional classification, in which entities are described by the intersection of two attributes at a time. The resulting matrix reveals the presence or absence and the nature of the entity at the intersection. An example of a paradigm is a kinship classification which can be organized into a grid, with gender (male/female) along one axis, and relation (parent, sibling, parent's sibling) along the other axis.

2.2.4. Faceted Classification

Facets will handle three or more dimensions of classification, where any complex entity could be viewed from a number of perspectives or facets (Denton 2003). The most widely used universal faceted classification schemes are Ranganathan's Colon Classification (Colon) and Bliss Bibliographic Classification – second edition (BC2). Colon Classification has five classic facets (dimensions) which are known as the PMEST (Personality, Matter, Energy, Space, and Time). Because these facets may not be enough in some classifications, BC2 contained thirteen

facets which have been found to be sufficient for the analysis of vocabulary in almost all areas of knowledge.

Faceted classifications are characterized by several properties: they do not require complete knowledge of the entities or their relationships; they are hospitable (can accommodate new entities easily); they are flexible; they are expressive; and they allow many different perspectives on the entities classified. On the other hand, the problems regarding faceted classifications are: the difficulty of choosing the right facets; and the lack of the ability to express the relationships between them. Table (2.1) presents a simplified example of how facets are used to classify a collection of socks (Broughton 2005):

Color	Pattern	Material	Function	Length
Black	Plain	Wool	Work	Ankle
Grey	Stripped	Polyester	Evening	Calf
Brown	Spotted	Cotton	Football	Knee
Green	Hooped	Silk	Hiking	
Blue	Checkered	Nylon	Protective	
Red	Novelty	Latex		

Table 2.1: Example of faceted classification (Broughton 2005, p.52)

2.2.5. User-Oriented Classification (Folksonomy):

Folksonomies are informal classifications that provide Web-specific classification issues. Folksonomy is a user-generated classification using freely chosen tags or keywords; that is retrieving the web content by using one's own descriptors. An important aspect in folksonomy is that it comprises terms in a flat namespace; that is, there is no hierarchy, and no directly specified 'type of" or 'part-of' relationships between these terms like the previously discussed formal classifications. It generates "related" tags automatically, which cluster tags based on common URLs. This is unlike formal classification schemes where there are "multiple kinds of explicit relationships between terms" (Uddin et al 2006).

Folksonomies require people to associate keywords with content and generate a list of popular keywords by tagging all the related documents that other users store in a particular website. Therefore folksonomies lead to a collection of web documents through user's choice of classification.

In contrast to formal classification techniques, this approach typically arises in non-hierarchical communities such as public websites and weblogs. (http://del.icio.us) is a one of those websites that use folksonomy classification, in which information is organized by its

primary users via assigning classifiers or tags. This approach is also known as free tagging or open tagging.

Advantages of classification schemes:

Websites that organize the information architecture with a classification scheme have the following advantages, which address many of the challenges described in (section 1.2):

- **Content browsing:** classification structures are helpful for the users unfamiliar with the content, structure, or terminology of a site.
- **Widening and narrowing searches:** hierarchical classification can be used to widen or narrow the search scope when required, which is useful for the huge amount of information on the web.
- **Multilingual access:** classification systems act as a switching language as they often use notations independent from a specific language, and therefore a searcher could enter search terms in a given language and those terms would then relate to the relevant parts of the classification system.
- **Flexible:** classification schemes are flexible and hospitable to accommodate new entities easily, which overcomes the problem of the web dynamicity.
- **Context-based:** classification scheme gives context to the search terms used, which overcomes the problem of homonyms (words with the same spelling but different meaning).
- **Machine-readable format:** many classification schemes are available in machine-readable form which ensures interoperability and overcomes the problem of web heterogeneity. For example, DDC is distributed by Machine Access Readable Catalogue (MARC).

2.3. TERM SIGNIFICANCE MEASURES:

Information retrieval systems are always based on retrieval techniques and ranking algorithms that require an indication about the terms significance. Term significance allows the retrieval system to rank relevant documents by giving an indication of the relevancy of a particular term to a certain document. Listed below are some of these significance measures:

2.3.1. Term Frequency:

The weight of a given term in a document, denoted by w_q, is simply the number of times, denoted by $N_{q,D}$, in which that term appears in the document it occurs within.

$$w_q = N_{q,D} \qquad \textit{Equation (2.1)}$$

2.3.2. Relative Term Frequency (RTF):

The Relative Term Frequency (RTF) is a variation of the Term Frequency. The RTF of a word is given by the ratio of the number of times a term appears in a web page to the frequency of the most frequent word in that web page ($N_{PageMax}$).

$$w_q = N_{q,D} / N_{PageMax} \qquad \textit{Equation (2.2)}$$

2.3.3. Paragraph Term Frequency:

This measure is similar to the RTF measure. In this case, the measure is restricted to a single paragraph only rather than the whole web page.

$$w_q = N_{q,p} / N_{Para\ Max} \qquad \textit{Equation (2.3)}$$

2.3.4. Word Emphasis Function:

This measure is mathematically difficult to calculate because of the various kinds of emphasis functions such as Bold, Italics, Underlining, Headings, Emphasized Text, Lists (Bullets) etc. In this case, HTML tags are used to constitute towards word emphasis in an HTML structured document.

2.3.5. Word Position:

It has been found that usually, the most important information in a text fragment is contained in the first third and the final third part. Thus, more weight is given to the initial third of the sentences that occur in a paragraph as well as the concluding third of the sentences in a paragraph.

2.3.6. Inverse Document Frequency (IDF):

Inverse document frequency (IDF) is a statistical measure of determining term significance (Spärck-Jones 1972; 2004). It attempts to capture how significant or insignificant a particular term is, by interpreting how many documents a term appears in relative to the document collection:

$$IDF_q = \log\left(\frac{N}{n_q}\right) \qquad \textit{Equation (2.4)}$$

N: number of documents in the collection

n_q: number of documents in which term q occurs.

This statistical measure is based on the notion that terms that occur in many documents in a document collection are more general and considered less important than terms that occur in fewer documents, which are considered more specific and thus more important. From a

retrieval perspective, this means that a specific term should provide a higher degree of potential relevancy to a document than a less specific term.

IDF is typically combined with Term Frequency (TF) to form the TFIDF measure, which has been confirmed as the best term weighting method.

2.3.7. Robertson-Spärck-Jones Weight:

Robertson-Spärck-Jones weight (RSJ) is another type of statistical term weighting scheme (Robertson & Sparck-Jones 1976). Similar to IDF, it also relies on observing how many documents a term appears within, but extends IDF by incorporating relevance information about terms that may be available. Given a query Q, the Robertson-Spärck-Jones weight for term q in Q is:

$$RSJ_q = \log_e \left[\frac{(r_q + 0.5)/(R - r_q + 0.5)}{(n_q - r_q + 0.5)/(N - n_q - R + r_q + 0.5)} \right] \qquad \textit{Equation (2.5)}$$

r_q: number of relevant documents containing q.

R: number of relevant documents known to be relevant for Q.

Although this weighting scheme has been used and proven to perform effectively, it has the disadvantage of having to obtain relevance judgments to be able to complete the formulation. Therefore, a simpler variation of the Robertson-Spark-Jones has been used, in which it does not base its formulation on the relevance information (r_q *and* R). This simplified variation is given by:

$$RSJ_q = \log_e \left[\frac{(N - n_q + 0.5)}{(n_q + 0.5)} \right] \qquad \textit{Equation (2.6)}$$

2.4. DOCUMENT RANKING TECHNIQUES:

Document ranking is the most important part of the information retrieval system. A ranking technique relies on the term weights in order to evaluate the similarity (relevance) between a submitted query (*Q)* and a document (*D)*. The following sections discuss some of these ranking techniques used in web information retrieval:

2.4.1. Inner Product

The similarity of a submitted query Q with potentially relevant documents can be calculated by the inner product of the query vector Q with the document vector D.

$$Sim_{Q,D} = \sum_{q \in Q} w_{q,Q} \times w_{q,D} \qquad \textit{Equation (2.7)}$$

$w_{q,Q}$: is the weight of term q in the query vector Q.

$w_{q,D}$: is the weight of term q in the document vector D.

Inner product forms the basis of TFIDF ranking which was stated in section 2.3.6 where IDF is used for $w_{q,Q}$ and TF is used for $w_{q,D}$. But it is not necessary to use the TFIDF weighting, since an inner product ranking can be used with other combinations of weights, in which $w_{q,Q}$ should give an indication of query term significance while $w_{q,D}$ should give an indication of document term significance.

Tables (2.2), (2.3) and (2.4) show how inner product calculation is performed. The inner product similarity of a document to a query is the weighted sum of the document term weights for all terms in the query that also occur in the document. Each document term weight is weighed by the corresponding query term and added to the similarity score.

	Term Weights (w_q)		**Similarity**
	q_1: car	***q_2: accident***	**(*Sim $_{Q,D}$*)**
Query (Q)	1	1	
Document 1 (D_1)	0.4	0.5	0.9
Document 2 (D_2)	0.5	0.4	0.9

Table 2.2: Inner Product Similarities – Scenario 1.

	Term Weights (w_t)		**Similarity**
	q_1: car	***q_2: accident***	**(*Sim $_{Q,D}$*)**
Query (Q)	0.8	1	
Document 1 (D_1)	0.4	0.5	0.82
Document 2 (D_2)	0.5	0.4	0.8

Table 2.3: Inner Product Similarities – Scenario 2.

	Term Weights (w_t)		**Similarity**
	q_1: car	***q_2: accident***	**(*Sim $_{Q,D}$*)**
Query (Q)	0.8	1	
Document 1 (D_1)	0.3	0.6	0.84
Document 2 (D_2)	0.6	0.4	0.88

Table 2.4: Inner Product Similarities – Scenario 3.

Term weights are assigned to query terms using a query term weighting technique such as TF, and term weights are assigned to documents using a document term weighting technique such as IDF. Term weights are between [0, 1] where a value close to one indicates high importance and a value of close to zero indicates low importance.

Comparing the above-mentioned scenarios, the impact of term weights on the inner product ranking is significant. A slight modification of term weights can have a significant influence on the ranking of retrieved documents and consequently affect the accuracy and effectiveness of a retrieval system.

2.4.2. Vector Model

The vector model is one of the popular classical information retrieval models that proposes a framework in which partial matching is possible. The vector model has become the most widely used information retrieval model because of its simplicity, ease of implementation, and effectiveness.

In vector model, both queries and documents are represented as vectors of non-binary weights, and then a similarity measure is used to determine the closeness between a document vector and a query vector. The vector model constructs multidimensional document vector representations for each document in the collection. This is established by generating a document-term matrix, in which each row represents a document and each column represents a single unique term. In its simplest form, the document-term matrix contains a count of the number of times a term occurs within a document. But various weighting schemes based on TFIDF have been proposed to go beyond this simple count scheme. In TFIDF, the more a term appears in a particular document and the less it appears in other documents in the document collection, the more important the term is at describing that document.

Once the weighted document-term matrix is constructed, a similarity measure, such as the cosine angle, is used to determine the closeness between a document vector and a query vector. The cosine angle measures the degree of relevance between a query and a document, and will be responsible for document ranking during retrieval. The cosine angle measures the similarity as follows:

$$Sim_{Q,D} = \left[\frac{\sum_{q} w_{q,Q} \times w_{q,D}}{\sqrt{\sum_{q} w_{q,Q}^2} \times \sqrt{\sum_{q} w_{q,D}^2}} \right] \qquad \textit{Equation (2.8)}$$

where $w_{q,D}$ is the weight of term q in the document vector D, while $w_{q,Q}$ is the weight of term q in the query vector Q.

Other similarity measures could be used with the vector model such as the inner product, pseudo-cosine, dice, and overlap measures. If the cosine angle is used, the document with the smallest cosine angle, when compared with the query vector, is ranked as the highest scoring document as shown in figure (2.3).

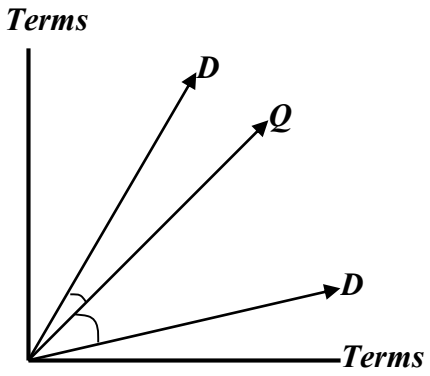

Figure 2.3: Cosine Angle of Vector Model

The extended vector model is an expansion of the classic vector model, in which multiple sub-vectors are used to represent a single document. Each sub-vector represents a different concept class of information. The model improves the effectiveness of retrieval by taking advantage of the information represented in the conceptual classes such as author's name, keywords, and bibliographic citations.

2.4.3. Probabilistic Model

By applying the probability theory to information retrieval, a document can be retrieved based on the probability that it is relevant to a submitted query. Having a set of queries, a set of documents, and a set of relevance judgments as training data, the weighted probability for each term in each document can be calculated. The weighted probability of each term reflects the probability that the document is relevant to the query, given that the term exists in the query as well.

The probabilistic model takes the assumption of that there is no dependency or relationship between the terms and each other. Although this assumption is unrealistic, it is important because the probability calculation of the term and query weights is based on this assumption. Ignoring this assumption means an enormous number of terms combinations which leads to an unrealistic incredible number of joint probability calculations, which is obviously infeasible.

Inference networks are considered as a probabilistic model, where evidential reasoning is used to determine whether a document is relevant to a query or not.

2.4.4. Okapi BM25

Okapi BM25 is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed by Stephen E. Robertson, Karen Spärck-Jones, and others. The name of the actual ranking function is BM25. To set the right context, however, it usually referred to as "Okapi BM25".

BM25, and its newer variants, e.g. BM25F (a version of BM25 that can take document structure and anchor text into account), represent state-of-the-art retrieval functions used in document retrieval, such as Web search.

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing terms $q_1,..,q_n$, the BM25 score of a document D is:

$$Sim_{(Q,D)} = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.\left(1-b+b.\frac{|D|}{avgdl}\right)} \qquad \textit{Equation (2.9)}$$

where $f(q_i,D)$ is q_i's term frequency in the document D, while |D| is the length of the document D (number of words), and $avgdl$ is the average document length in the text collection from which documents are drawn; k_1 and b are free parameters, usually chosen as $k_1 = 1.2$ and $b = 0.75$. $IDF(q_i)$ is the inverse document frequency weight of the query term q_i. It is usually computed using equation (2.6) in section (2.3.7).

The IDF component is where the probabilistic nature of BM25 becomes apparent. Suppose a query term q appears in $n(q)$ documents. Then a randomly picked document D will contain the term with a probability of

$$\left(\frac{n(q)}{N}\right) \qquad \textit{Equation (2.10)}$$

Therefore, the information content of the message "*D contains q*" is:

$$-\log\frac{n(q)}{N} = \log\frac{N}{n(q)} \qquad \textit{Equation (2.11)}$$

Now suppose we have two query terms q_1 and q_2. If the two terms occur in documents entirely independently of each other, then the probability of seeing both q_1 and q_2 in a randomly picked document D is:

$$\left(\frac{n(q_1)}{N}\right).\left(\frac{n(q_2)}{N}\right) \qquad \textit{Equation (2.12)}$$

and the information content of such an event is:

$$\sum_{i=1}^{2} log\ \frac{N}{n(q_i)} \qquad \textit{Equation (2.13)}$$

2.4.5. Fuzzy Logic

The fuzzy set theory, or fuzzy logic, has been applied to information retrieval. Document representations and queries are usually imprecise; therefore it is difficult to calculate the importance of a term as a descriptor. Fuzzy logic provides a framework in an attempt to deal with the unique imprecise nature of information on the web.

The Fuzzy Information Retrieval System (FIRST) implemented a knowledge-based information retrieval system using fuzzy logic (Lucarella & Morara 1991). A concept network is used as a knowledge base whose links represent relationships between concepts and documents. Each link between a concept and a document has an associated weight assigned by a membership function. Once the query is submitted, the concept network is activated and a number of fuzzy inference rules are applied, which retrieves a list of documents considered relevant to the query.

Ogawa et al, as cited by Zakos (2005), proposed a different approach of incorporating fuzzy logic into information retrieval. A term-to-term similarity matrix is used to convert crisp (hard) terms in a document representation into fuzzy document representation. Then, a fuzzy retrieval algorithm computes the relevance of documents to a query that is subdivided into sub-queries; the query is subdivided according to the logical operators that are used, such as (AND), (OR) and (NOT). Each sub-query is compared to each document to determine its relevance, creating a fuzzy set for each sub-query. The fuzzy retrieval algorithm computes the overall relevance based on the intersection or union of two fuzzy sets according to the logical operators. This approach also uses a learning technique that updates the weights of the term-to-term matrix according to relevance feedback from the user.

2.4.6. Hyperlink Analysis

The web is a collection of hyperlinked web documents that contain pointers to each other. As broad queries may result in thousands or millions of documents, hyperlink information plays a good role in deciding which of these documents are of good quality.

Link analysis has been used successfully for deciding which web pages to add to the collection of documents (i.e., which pages to crawl), and how to order the documents matching a user query (i.e., how to rank pages). It has also been used to categorize web pages, to find pages that are related to given pages, to find duplicated web sites, and various other problems related to web information retrieval (Henzinger 2000).

A hyperlink is a reference of a web page that is contained in a web page (A). When the hyperlink is clicked on in a web browser, the browser displays page (B). Thus, links are usually either navigational aids that, for example, bring the reader back to the homepage of the site, or links that point to pages whose content augments the content of the current page. The second kind of links tends to point to high-quality pages that might be on the same topic as the page containing the link.

Hyperlink information is either based on a set of documents during retrieval (e.g. HITS) or the analysis of the whole document collection during indexing (e.g. PageRank).

The algorithms of (Kleinberg 1999) illustrate how hyperlink information is useful in web search when using a set of retrieved documents. These algorithms are based on the concept that if a page (A) points to page (B) then *A* has a kind of conferred authority on *B*. Therefore, the more pages that point to (B), the more (B) is considered authoritative on the topic it represents; in other words, (B) is considered as an authority of information to that topic. On the other hand, a hub is the page that has multiple links to authoritative pages. Thus, a good hub is the one that points to many good authoritative pages, and a good authority is the one pointed to by many good hubs. The relationship between hubs and authorities is shown in figure (2.4). Therefore, this mutually reinforcing relationship between hubs and authorities can be used to determine the quality of pages during the ranking of documents, by calculating the scores of hubs and authorities. This method is known as HITS.

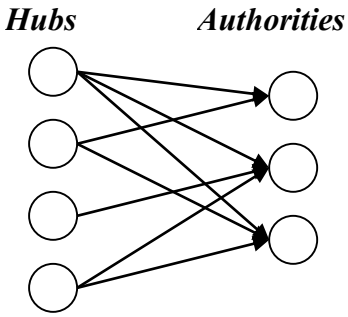

Figure 2.4: Relationship between hubs and authorities

Unlike Kleinberg's method which deals only with hyperlinks between a retrieved set of documents, another method of interpreting hyperlinks is based on a global analysis of the entire document collection. PageRank is an algorithm that calculates document scores by considering all the links in the entire document collection (Brin & Page 1998). It uses link information to model user's behavior by evaluating the probability that the user will visit a certain webpage. This probability or PageRank of a page is used in ranking it during retrieval.

Also, a modified version of PageRank has been proposed by factoring the impact of pages undiscovered by crawlers as well as the old pages that are no longer maintained (Eiron et al 2004).

2.5. RETRIEVAL ACCURACY MEASURES:

"*Precision*" and "*Recall*" are two related methods that are used to measure the retrieval accuracy of an information retrieval system with respect to a given query. Each submitted query has a number of associated relevant documents that are considered as correct answers for the query as shown in figure 2.5:

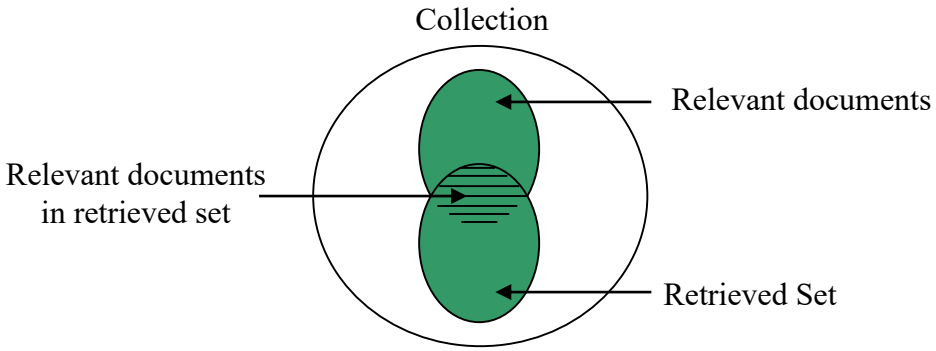

Figure 2.5: Relevant documents in a retrieved set for a given query

For a given query, these two accuracy measures are evaluated as follows:

a) $$Precision = \frac{Number\ of\ relevant\ documents\ in\ retrieved\ set}{Number\ of\ retrieved\ documents}$$ ***Equation (2.14)***

b) $$Recall = \frac{Number\ of\ relevant\ documents\ in\ retrieved\ set}{Number\ of\ relevant\ documents}$$ ***Equation (2.15)***

Precision is the measure of relevancy of the retrieved list relative to the number of documents that appear in the list. So, if only one document is retrieved and that document is considered to be relevant, then the precision would be 100% (1/1). But if 100 documents were retrieved and only ten of these documents were considered relevant, then the precision would be measured at 10% (10/100).

Unlike precision, recall does not take into consideration the number of retrieved documents. Recall is a measure of the relevancy of the retrieved list relative to the total number of relevant documents in the document collection. Consider a query that is associated with 20 relevant documents in a document collection of 100 documents. If this query is submitted to a system that retrieves only 10 relevant documents, then the recall would be measured at 50% (10/20).

The ultimate goal of an information retrieval system is to perform with 100% precision but also with 100% recall. But, there is an inverse relationship between precision and recall in which the information retrieval system will record a high measure of precision at a low recall rate.

To evaluate the retrieval accuracy of an algorithm over all test queries, the precision is averaged at each recall level as follows:

$$\overline{P}(r) = \frac{\sum_{i=1}^{N_q} P_i(r)}{N_q}$$ ***Equation (2.16)***

where $\overline{P}(r)$ is the average precision at recall level r, N_q is the number of queries used, and $P_i(r)$ is the precision at recall level r for the i^{th} query.

2.6. CONCEPT-BASED RETRIEVAL:

Concept-based retrieval does not refer to a strict information retrieval model. Many approaches attempt to improve information retrieval by incorporating the semantics of words into a retrieval model. Some of these approaches use a knowledge base and others do not.

The WordNet ontology is a large lexical English database whose structure makes it a useful tool for computational linguistics and natural language processing. WordNet classifies

the four parts of speech POS (nouns, verbs, adjectives and adverbs) into synonymous sets (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations such as, Synonyms, Hypernyms, Hyponyms, Troponyms, etc. (Fellbaum et al 1990; Miller et al 1990; Gross et al 1990).

- Synonyms: *(... have the same meaning as "q")*
- Antonyms: (… *are opposites to "q"*)
- Noun Hypernyms: *("q" is a kind of ...)*
- Hyponyms: *(... are kinds of "q")*
- Verb Hypernyms: *("q" is one way to ...)*
- Troponyms: *(... are particular ways to "q")*
- Holonyms: *("q" is a part of ...)*
- Meronyms: *(...are parts of "q")*

Every concept in WordNet can be traced up to a root concept called a unique beginner, which represents the most top level abstract level in the hierarchy. Table (2.5) shows the number of words and synsets in different parts of speech, presented in the WordNet documentation:

Part of Speech	Unique Beginners	Synsets
Noun	117,097 words	81,426 synsets
Verb	11,488 words	13,650 synsets
Adjective	22,141 words	18,877 synsets
Adverb	4,601 words	3,644 synsets
Total	**155,327 words**	**117,597 synsets**

Table 2.5: Number of words and synsets in WordNet 2.1

However, *WordNet* can sometimes cause problems because it is very specific in its definitions. For example, the word "*brick*" has two senses; the first one, which is commonly used, refers to the "brick" as "rectangular block of clay baked by the sun or in a kiln; used as a building or paving material", while the second sense, which is rarely used, refers to it as "a good fellow; helpful and trustworthy".

2.7. INFORMATION PERSONALIZATION:

As the number of web pages increases dramatically, inexperienced users feel that they are looking for a needle in this growing haystack. To address this problem, personalization becomes a popular remedy to customize the Web environment towards a user's preference.

Generally, most modern search engines do not return personalized results. That is, the result of a search for a given query is identical, independent of the user submitting the query. Hence, by ignoring the user's preferences during the search process, the search engines may return a large amount of irrelevance data (Shahabi & Chen 2003). For example, a geographer and a programmer may use the same word “java”. By this query, some users may mean Java Indonesian Island, while other users may be interested in Java Programming Language. Moreover, a user’s information needs may change over time. The same user may use “Java” sometimes to mean the Indonesian Island and some other times to mean the Programming Language.

For a given query, a personalized search can be implemented on either the server side (search engine) or the client side (user’s computer). Personalized search implemented on the server side raises privacy concerns when information about users is stored on the server. A personalized search on the client side can be achieved by query expansion and/or result processing. By adding extra query terms associated with user interests or search context, the query expansion approach can retrieve different sets of results. The result processing includes result filtering, such as removal of some results, and reorganizing, such as re-ranking, clustering, and categorizing the results.

Other studies showed how to exploit implicit user modeling to intelligently personalize information retrieval and improve search accuracy (Shen et al 2005). They emphasized the use of immediate search context and implicit feedback information as well as eager updating of search results to maximally benefit a user.

Moreover, personalized spiders (crawlers) were also used for web search and analysis (Chau et al 2001). Two systems, namely CI Spider and Meta Spider, have been built based on a client-based architecture that incorporates noun phrasing and self-organizing map techniques.

2.8. CROSS-LANGUAGE RETRIEVAL TECHNIQUES:

The web is essentially multilingual. Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query, in order to cross the language barriers. CLIR has many useful applications. For example, multilingual searchers might want to issue a single query to a multilingual collection, or searchers with a limited active vocabulary, but good reading comprehension in a second language such as English, might prefer to issue queries in their most fluent language such as Arabic (Youssef 2001).

A 2004 survey of 2,024 million web pages determined that by far the most web content was in English (56.4%) and (43.6%) was in non-English (Sigurbjornsson et al 2005). Tables (2.6) and (2.7) show these statistics of the web contents and web population, respectively, according to language.

Web content by language		
Language	**Internet Pages (in millions)**	**Web Content (percentage)**
English	1142.5	56.4
Non-English	882.2	43.6
Euro-non-English	536.9	26.5
Dutch	38.8	1.9
French	113.1	5.6
German	156.2	7.7
Italian	41.1	2.0
Polish	14.8	0.7
Portuguese	29.4	1.5
Russian	33.7	1.7
Scandinavian	17.4	1.3
Spanish	59.9	3.0
Other European	32.5	1.1

Table 2.6: Web content by language

Web population by language		
Language	**Internet Access (in millions)**	**Web Population**
English	295.4	35.2
Non-English	544.5	64.8
Euro-non-English	285.5	35.7
Dutch	14.0	1.7
French	33.9	4.2
German	55.3	6.9
Italian	30.4	3.3
Polish	9.6	1.2
Portuguese	24.4	3.1
Russian	6.5	0.8
Scandinavian	12.8	1.6
Spanish	72.0	9.0
Other European	26.6	3.9

Table 2.7: Web population by language

Figure (2.6) is a graphical representation for tables (2.6) and (2.7), in which it illustrates the language gap between the amount of contents and the number of users. Therefore, researchers of CLIR seek to support the process of finding documents written in one natural language with automated systems that can accept queries expressed in other languages.

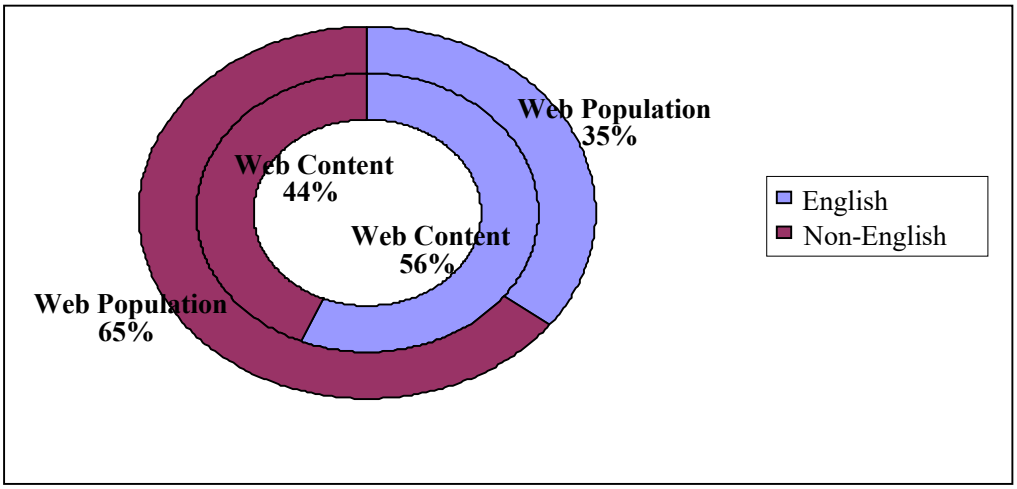

Figure 2.6: Web Content vs. Web Population (Sep 2004 statistics)

In the workshops of the Cross Language Evaluation Forum (CLEF) from year 2000 till present, authors and researchers presented a number of approaches, in order to bridge the language gap, or in other words, cross the language barriers.

In CLIR, either documents or queries are translated. There are three main approaches to CLIR: machine-readable dictionary, machine translation, and comparable or parallel corpora. Appendix A shows a list of the most widely used systems for each of the three approaches.

2.8.1. Machine-Readable Dictionaries (MRD)

Dictionary-based methods perform query translation by looking-up terms on a bilingual dictionary and building a target language query by adding some or all of the translations. In the MRD, there are several lookup techniques, such as the Every-Match, the First-Match and the Two-Phase Method (Aljlayl & Frieder 2001).

Although, dictionary-based method yields to ambiguous translations, the practicality of dictionary-based translation is increasing due to the greater availability of machine-readable bilingual dictionaries. Moreover, a domain-specific dictionary (e.g. medical, military, business, etc.) reduces the ambiguity compared to general dictionaries.

Several methods were developed using MRDs for Spanish-English CLIR (Ballesteros & Croft 1997). The first experiment was designed to test the effect of word-by-word translation on retrieval performance. The average precision dropped 50-60%; the reason behind the low effectiveness is that many noise terms were added. To improve the effectiveness, they introduced the notion of pre-translation and post-translation methods. Another experiment investigated the effect of phrasal translation in improving effectiveness.

2.8.2. Machine Translation (MT)

Machine translation systems can be defined as any computer-based process to transform a text from one language to another. The basic task of any machine translation system is to analyze the source text, including morphological, syntactic, and semantic analysis using special purpose lexicons, and target language generation. There are two basic approaches to MT: translating the documents or translating the queries. But, usually the query is to be translated into the language of the documents, and not the opposite.

Many authors criticize the MT-based method due to the fact that the current translation quality is poor. A study compared the retrieval effectiveness of French-English CLIR using SYSTRAN machine translation system with the effectiveness of their EMIR dictionary-based query translation; the experimental results showed that the EMIR was more effective than MT-based technique using SYSTRAN (Radwan & Fluhr 1993). Other researchers, in contrast,

showed that machine translation approaches could achieve reasonable effectiveness. Participants in the Text Retrieval Conference (TREC-8) concluded that MT-based CLIR is an effective strategy.

2.8.3. Comparable and Parallel Corpora

The term “corpora”, a Spanish word that means “bodies”, refers to web documents. In corpus-based methods, queries are translated on the basis of the terms that are extracted from parallel or comparable document collections. Using dictionary-based cross language retrieval dictionaries alone in CLIR is problematic; some of the translation alternatives of a word may differ from the meaning intended by the user.

In corpus-based methods, translation knowledge is derived from multilingual text collections using various statistical methods. Such collections can be aligned or unaligned. In aligned multilingual collections, each source language document is mapped to a target language document. If the paired documents are exact translations of each other, the collection is a parallel corpus. Document alignments can also be used to disambiguate dictionary-based query translation. Usually this works as follows. A source language query is first translated with a machine-readable dictionary. If multiple translation alternatives occur, the original query is run against the source language documents of the aligned collection.

Comparable corpora consist of document pairs that are not translations of each other but share similar topics.

- **Parallel-corpus-based approaches:**

Collecting parallel texts of different language versions from the web has recently received much attention. The BBN research team obtained a collection of documents from the United Nations that included translation-equivalent document pairs in English and Arabic. Word-level alignments were created using statistical techniques and then used as basis for determining frequently observed translation pairs. Davis and Dunning (1995) used a Spanish-English parallel corpus and evolutionary programming for query translation. Landauer (1994) introduced another method for which no query translation is required. This method is called Cross-Language Latent Semantic Indexing (CL-LSI), and requires a parallel corpus.

- **Comparable-corpus-based approaches:**

Unlike parallel corpora, comparable corpora are collections of texts from pairs or multiples of languages, which can be contrasted because of their common features, in the topic, the domain, the authors or the time period. This property made comparable corpora more

abundant, less expensive and more accessible through the World Wide Web (Talvensaari et al 2007).

CHAPTER 3

SYSTEM ARCHITECTURE.

This chapter presents the techniques that are used to form the proposed approach in this research. Section (3.1) gives an overview of the proposed approach and describes each component of the system. Section (3.2) presents the concept-based term weighting technique, while section (3.3) describes the context matching technique. This is followed by section (3.4) which presents a case study to illustrate the effectiveness of context matching. Section (3.5) describes the document ranking techniques, and section (3.6) presents the parallel corpus construction technique. Finally, section (3.7) presents a summary of the system architecture.

CHAPTER 4

EXPERIMENTAL RESULTS.

This chapter provides the experimental results of the techniques presented in the system architecture at the previous chapter. The techniques that are experimented are Concept-based Term Weighting, Context Matching, Document Ranking, and Parallel Corpus Construction. The first section describes the experimental environment, while following sections test each technique and provide the corresponding experimental results.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This chapter provides a conclusion for the techniques used for retrieval, ranking and translation. Section (5.1) presents a conclusion for each of the concept-based term weighting, context matching, weighted document ranking, and parallel corpus construction. This is followed by section (5.2) which provides an outline for the future work that could be done to optimize, support or substitute any component of the presented techniques.

APPENDIX A

MT, MRD, CORPORA

Table A.1: Popular Machine Translation Systems, Machine Readable Dictionaries, and Corpora:

Machine Translation (MT)	Babel Fish (Systran System)
	Google.com
	FreeTranslation.com
	InterTran
	Reverso.net (in French)
	Reverso Online (in English)
Machine-Readable Dictionaries (MRD)	Freedict.com
	Foreignword.com (Babylon)
	Leo Dictionaries
	YourDictionary.com
	Other Dictionaries
Corpora	Linguistic Data Consortium (LDC)
	European Language Resource Association (ELRA)
	EuroWordNet (license available from ELRA)
	Part-of-speech tagger (English) – from University of Edinburgh (UK)
	Other CLIR resources from University of Maryland at College Park
	CJK linguistics resources (for Asian languages)
	Canadian Hansard Corpus

APPENDIX B

PORTER STEMMING ALGORITHM

B.1 Introduction

The Porter-Stemming algorithm, written and maintained by Martin Porter (2006), is used for stripping suffixes on English language words. The Porter-Stemmer algorithm removes suffixes by automatic means, which is essential in the area of Information Retrieval. In a typical Information Retrieval environment, we have a large collection of text in the form of documents with each document described by special sections such as titles, headings, abstracts etc. Ignoring the origin of specific words, every document consists of a vector of terms i.e. a stem with a particular suffix. All the terms with a common stem usually have the same meaning, such as:

CONNECT, CONNECTED, CONNECTING, CONNECTION, CONNECTIONS.

It has been found that the performance of any Information Retrieval system if the terms having the common stems are clubbed together. This may be achieved by removing various suffixes such as – ED, -ING, -ION, etc. for example in the above case stripping suffixes would leave us with the stem CONNECT. In addition to removing suffixes, it also reduces the total number of terms in an Information Retrieval system, thus reducing the total size and complexity and size of the system.

B.2 Definitions

Before presenting the Porter-Stemmer algorithm, we need to define some terms:

Definition: A **consonant** in a word is a letter other than *A, E, I, O or U* and other than *Y* preceded by a consonant.

Definition: If a letter is NOT a consonant then it is said to be a **vowel**

A consonant will be denoted by ‘c’, a vowel by ‘v’. A list ccc… of length greater than zero will be denoted by C while a list vvv… of length greater than zero will be denoted by V. Any word, or a part of a word, thus has to be one of the four forms,

CVCV … C

CVCV … V

VCVC … C

VCVC … V

These may all be represented by the single form:

[C] VCVC ... [V]

where, the square brackets denote arbitrary presence of their contents.

Using (VC){m} to denote VC repeated m times, this may again be written as:

[C](VC){m}[V].

Definition: m will be called the **measure** of any word or word part when represented in this form.

The case m = 0 covers the null word. Here are some examples:

m=0 TR, EE, TREE, Y, BY.

m=1 TROUBLE, OATS, TREES, IVY.

m=2 TROUBLES, PRIVATE, OATEN, ORRERY.

The *rules* for removing a suffix will be given in the form:

(condition) S1 → S2;

This means that if a word ends with the suffix S1, and the stem before S1 satisfies the given condition, S1 is replaced by S2. The condition is usually given in terms of m, e.g.

(m > 1) EMENT →;

Here S1 is 'EMENT' and S2 is null. This would map REPLACEMENT to REPLAC, since REPLAC is a word part for which m = 2. The 'condition' part may also contain the following:

*S - the stem ends with S (and similarly for the other letters).

v - the stem contains a vowel.

*d - the stem ends with a double consonant (e.g. -TT, -SS).

*o - the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

And the condition part may also contain expressions with AND, OR or NOT so that,

(m>1 and (*S or *T)) tests for a stem with m>1 ending in S or T; while (*d and not (*L or *S or *Z)) tests for a stem ending with a double consonant other than L, S or Z.

Elaborate conditions like this are required only rarely.

In a set of rules written beneath each other, only one is obeyed, and this will be the one with the longest matching S1 for the given word. For example, with

SSES → SS

IES → I

SS → SS

S → (here the conditions are all null) CARESSES maps to CARESS since SSES is the longest match for S1. Equally, CARESS maps to CARESS (S1='SS') and CARES to CARE (S1='S'). In the rules below, examples of their application, successful or otherwise, are given on the right in lower case.

B.3 Algorithm

The algorithm consists of five steps. Below listed is each step in detail with the rules that it contains.

Step 1-A

SSES → SS; caresses → caress

IES → I; ponies → poni, ties → ti

SS → SS; caress → caress

S →; cats → cat

Step 1-B

(m>0) EED → EE; feed → feed, agreed → agree

(*v*) ED →; plastered → plaster, bled → bled

(*v*) ING →; motoring → motor, sing → sing

If the second or third of the rules in Step 1b is successful, the following is done:

AT → ATE; conflat(ed) → conflate

BL →BLE; troubl(ed) → trouble

IZ → IZE; siz(ed) → size

(*d and not (*L or *S or *Z)) → single letter; hopp(ing) → hop, tann(ed) → tan,

; fall(ing) → fall, hiss(ing) → hiss

; fizz(ed) → fizz,

(m=1 and *o) → E; fail(ing) → fail, fil(ing) → file

The rule to map to a single letter causes the removal of one of the double letter pair. The -E is put back on -AT, -BL and -IZ, so that the suffixes -ATE, -BLE and -IZE can be recognized later. This E may be removed in step 4.

Step 1-C

(*v*) Y → I; happy → happi, sky → sky

Step 1 deals with plurals and past participles. The subsequent steps are much more straightforward.

Step 2

(m>0) ATIONAL → ATE; relational → relate

(m>0) TIONAL → TION; conditional → condition

(m>0) ENCI → ENCE; valenci → valence

(m>0) ANCI → ANCE; hesitanci → hesitance

(m>0) IZER → IZE; digitizer → digitize

(m>0) BLI → BLE; possibli → possible

(m>0) ALLI → AL; radically → radical

(m>0) ENTLI → ENT; differentli → different

(m>0) ELI →E; vileli → vile

(m>0) OUSLI → OUS; analogousli → analogous

(m>0) IZATION → IZE; vietnamization → vietnamize

(m>0) ATION → ATE; predication → predicate

(m>0) ATOR → ATE; operator → operate

(m>0) ALISM → AL; feudalism → feudal

(m>0) IVENESS → IVE; decisiveness → decisive

(m>0) FULNESS → FUL; hopefulness → hopeful

(m>0) OUSNESS → OUS; callousness → callous

(m>0) ALITI → AL; formaliti → formal

(m>0) IVITI → IVE; sensitiviti → sensitive

(m>0) BILITI → BLE; sensibiliti → sensible

(m>0) logi → log; archaeologi → archaeolog

The test for the string S1 can be made fast by doing a program switch on the penultimate letter of the word being tested. This gives an even breakdown of the possible values of the string S1. It will be seen in fact that the S1-strings in step 2 are presented here in the alphabetical order of their penultimate letter. Similar techniques may be applied in the other steps.

Step 3

(m>0) ICATE → C; triplicate → triplic

(m>0) ATIVE →; formative → form

(m>0) ALIZE → AL; formalize → formal

(m>0) ICITI → IC; electriciti → electric

(m>0) ICAL → IC; electrical → electric

(m>0) FUL →; hopeful → hope

(m>0) NESS →; goodness → good

Step 4

(m>1) AL →; revival → reviv

(m>1) ANCE →; allowance → allow

(m>1) ENCE →; inference → infer

(m>1) ER →; airliner → airlin

(m>1) IC →; gyroscopic → gyroscop

(m>1) ABLE →; adjustable → adjust

(m>1) IBLE →; defensible → defens

(m>1) ANT →; irritant → irrit

(m>1) EMENT →; replacement → replac

(m>1) MENT →; adjustment → adjust

(m>1) ENT →; dependent → depend

(m>1 and (*S or *T)) ION →; adoption → adopt

(m>1) OU →; homologou → homolog

(m>1) ISM →; communism → commun

(m>1) ATE →; activate → activ

(m>1) ITI →; angulariti → angular

(m>1) OUS →; homologous → homolog

(m>1) IVE →; effective → effect

(m>1) IZE →; bowdlerize → bowdler

The suffixes are now removed. All that remains is a little tidying up.

Step 5-A

(m>1) E →; probate → probat, rate → rate

(m=1 and not *o) E →; cease → ceas

Step 5-B

(m > 1 and *d and *L) → single letter ; controll → control, roll → roll

The algorithm is careful not to remove a suffix when the stem is too short, the length of the stem being given by its measure, m. There is no linguistic basis for this approach. It was merely observed that m could be used quite effectively to help decide whether it was wise to take off a suffix. For example, in the following two lists:

List A	List B
RELATE	DERIVATE
PROBATE	ACTIVATE
CONFLATE	DEMONSTRATE
PIRATE	NECESSITATE
PRELATE	RENOVATE

-ATE is removed from the list B words, but not from the list A words. This means that the pairs (DERIVATE / DERIVE), (ACTIVATE / ACTIVE), (DEMONSTRATE / DEMONSTRABLE), (NECESSITATE / NECESSITOUS), will conflate together. The fact that no attempt is made to identify prefixes can make the results look rather inconsistent. Thus, PRELATE does not lose the -ATE, but ARCHPRELATE becomes ARCHPREL. In practice, this does not matter too much, because the presence of the prefix decreases the probability of an erroneous conflation.

Complex suffixes are removed bit by bit in the different steps. Thus GENERALIZATIONS is stripped to GENERALIZATION (Step 1), then to GENERALIZE (Step 2), then to GENERAL (Step 3), and then to GENER (Step 4). OSCILLATORS is stripped to OSCILLATOR (Step 1), then to OSCILLATE (Step 2), then to OSCILL (Step 4), and then to OSCIL (Step 5).

APPENDIX C

WT2G TOPICS

TABLE C.1: WT2G TOPICS (401-450):

No.	Title	Description (Tokenized)
401	foreign minorities, Germany	language, cultural, differences, impede, integration, foreign, minorities, Germany
402	behavioral genetics	happening, field, behavioral, genetics, study, relative, influence, genetic, environmental, factors, individual's, behavior, personality
403	osteoporosis	information, effects, dietary, intakes, potassium. magnesium, fruits, vegetables, determinants, bone, mineral, density, elderly, men, women, preventing, osteoporosis, bone, decay
404	Ireland, peace talks	often, peace, talks, Ireland, delayed, disrupted, result, acts, violence
405	cosmic events	unexpected, unexplained, cosmic, events, celestial, phenomena, radiation, supernova, outbursts, new, comets, detected
406	Parkinson's disease	being, done, treat, symptoms, Parkinson's, disease, keep, patient, functional, long, possible
407	poaching, wildlife preserves	impact, poaching, world's, various, wildlife, preserves
408	tropical storms	tropical, storms, hurricanes, typhoons, caused, significant, property, damage, loss, life
409	legal, Pan Am, 103 legal	legal, actions, resulted, destruction, pan am, flight, 103, Lockerbie, Scotland, December 21 1988
410	Schengen agreement	involved, Schengen, agreement, eliminate, border, controls, western, Europe, hope, accomplish
411	salvaging, shipwreck, treasure	find, information, shipwreck, salvaging, recovery, attempted, recovery, treasure, sunken, ships

412	airport security	security, measures, effect, proposed, go, effect, airports
413	steel production	new, methods, producing, steel
414	Cuba sugar exports	sugar, Cuba, export, countries, import
415	drugs, Golden Triangle	drugs, known, trafficking, golden, triangle, area, burna, Thailand, Laos, meet
416	Three Gorges Project	status, three, gorges, project
417	creativity	find, ways, measuring, creativity
418	quilts, income	ways, quilts, used, generate, income
419	recycle, automobile tires	new, uses, developed, old, automobile, tires, means, tire, recycling
420	carbon monoxide poisoning	widespread, carbon, monoxide, global, scale
421	industrial waste disposal	disposal, industrial, waste, being, accomplished, industrial, management, world
422	art, stolen, forged	incidents, stolen, forged, art
423	Milosevic, Mirjana Markovic	find, references, Milosevic's, wife, Mirjana, Markovic
424	suicides	give, examples, alleged, suicides, aroused, suspicion, death, actually, being, murder
425	counterfeiting money	counterfeiting, money, being, done, modern, times
426	law enforcement, dogs	provide, information, use, dogs, worldwide, law, enforcement, purposes
427	UV damage eyes	find, documents, discuss, damage, ultraviolet, UV, light, sun, eyes
428	declining birth rates	countries, U.S., china, declining, birth, rate
429	Legionnaires' disease	identify, outbreaks, legionnaires', disease
430	killer bee attacks	identify, instances, attacks, humans, Africanized, killer, bees
431	robotic technology	latest, developments, robotic, technology
432	profiling, motorists, police	police, departments, use, profiling, stop, motorists
433	Greek, philosophy, stoicism	contemporary, interest, Greek, philosophy, stoicism
434	Estonia, economy	state, economy, Estonia

435	curbing population growth	measures, taken, worldwide, countries, effective, curbing, population, growth
436	railway accidents	causes, railway, accidents, world
437	deregulation, gas, electric	experience, residential, utility, customers, following, deregulation, gas, electric
438	tourism, increase	countries, experiencing, increase, tourism
439	inventions, scientific discoveries	new, inventions, scientific, discoveries, made
440	child labor	steps, taken, governments, corporations, eliminate, abuse, child, labor
441	Lyme disease	prevent, treat, Lyme, disease
442	heroic acts	find, accounts, selfless, heroic, acts, individuals, small, groups, benefit, others, cause
443	U.S., investment, Africa	extent, U.S., government, private, investment, sub-Saharan, Africa
444	supercritical fluids	potential, uses, supercritical, fluids, environmental, protection, measure
445	women clergy	countries, United, states, considering, approved, women, clergy, persons
446	tourists, violence	tourists, likely, subjected, acts, violence, causing, bodily, harm, death
447	Stirling engine	new, developments, applications, stirling, engine
448	ship losses	identify, instances, weather, main, contributing, factor, loss, ship, sea
449	antibiotics ineffectiveness	caused, current, ineffectiveness, antibiotics, against, infections, prognosis, new, drugs
450	King Hussein, peace	significant, figure, years, late, Jordanian, king, Hussein, furthering, peace, middle, east

APPENDIX D

QUERY EXPANDED TERMS

TABLE D.1: TOP 10 EXPANDED TERMS FOR ORIGINAL QUERY TERMS:

No.	Original Terms	Expanded Terms
401	foreign minorities Germany	"alien", "asylum", "foreign", "German", "Germany", "illegal", "immigration", "ins", "migrant", "minor"
402	behavioral genetics	"abuse", "BBP", "behavior", "gene", "genetic", "heritage", "homosexual", "psychiatry", "trait", "twin"
403	osteoporosis	"bone", "calcium", "estrogen", "flash", "hysterectomy", "menopausal", "osteoporosis", "progesterone", "uterine", "women"
404	Ireland peace talks	"bomb", "ceasefire", "Fein", "IRA", "Ireland", "nationalist", "peace", "talk", "unionist"
405	cosmic events	"Ashtar", "cosmic", "Cosmo", "detector", "Dior", "earth", "energy", "event", "particle", "ray"
406	Parkinson's disease	"Alzheimer", "brain", "cell", "disease", "dopamine", "dosage", "levodopa", "Parkinson", "patient", "receptor"
407	poaching wildlife preserves	"conserve", "habitat", "poach", "preserve", "rhino", "species", "Sumatran", "tiger", "wildlife", "WWF"
408	tropical storms	"cyclone", "hurricane", "precipitant", "radar", "storm", "thunderstorm", "tornado", "tropic", "weather", "wind"
409	legal Pan Am 103 legal	"103", "am", "bomb", "court", "dispute", "internal", "legal", "Libya", "Libyan", "Lockerby"
410	Schengen agreement	"agreement", "cannabis", "coffee", "drug", "Dutch", "lancet", "Netherlands", "pot", "Schengen"
411	salvaging shipwreck treasure	"dive", "expedient", "Indian", "Nicholas", "Russian", "salvaging", "sea", "ship", "shipwreck", "treasure"
412	airport security	"AAAE", "airline", "airport", "bomb", "cargo", "contract", "enhance", "FAA", "passenger", "secure"
413	steel production	"alloy", "furnace", "ingot", "metal", "production", "recycle", "schedule", "scrap", "steel", "steelmaking"
414	Cuba sugar exports	"Caribbean", "Cuba", "Cuban", "dollar", "economy", "export", "latin", "Spain", "Spanish", "sugar"

415	drugs Golden Triangle	"cocaine", "Colombian", "drug", "golden", "Gritz", "heroin", "opium", "triangle", "smuggle", "trafficker",
416	Three Gorges Project	"ADB", "Asia", "dam", "EGAT", "electric", "gorges", "hydro", "mw", "power", "project"
417	creativity	"Amazon", "auditorium", "creativity", "idea", "ISP" "Issar", "Neal", "photography", "seminar", "sponsor"
418	quilts income	"aid", "Falzarano", "family", "gay", "homosexual", "income", "knight", "mill" "PFOX", "quilt"
419	recycle automobile tires	"automobile", "car", "compressor", "dealer", "evaporate", "fee", "recycle", "retread", "scrap", "tire"
420	carbon monoxide poisoning	"appliance", "carbon", "detector", "gas", "heater", "monoxide", "oxygen", "poison", "smoke", "tobacco"
421	industrial waste disposal	"disposal", "hazard", "industrial", "injection", "landfill", "radioactive", "solid", "Texas", "waste", "wood"
422	art stolen forged	"500", "art", "forged", "forgery", "Indonesia", "Indonesian", "Jakarta", "museum", "stolen", "theft"
423	Milosevic Mirjana Markovic	"Belgrade", "Croatia", "Markov", :Milosevic", "Mirjana", "Serbia", "Serbian", "Srpska", "Yugoslavia"
424	suicides	"attempt", "depression", "emotion", "feeling", "ideate", "ill", "Japanese", "patient", "psychiatry", "suicide"
425	counterfeiting money	"bank", "card", "Chinese", "counterfeit", "dollar", "Fed", "fraud", "money", "tariff", "treasury"
426	law enforcement dogs	"1936", "1937", "dog", "enforcement", "FBI", "law", "Marihuana", "Marijuana", "Marshal", "weaver"
427	UV damage eyes	"damage", "depletion", "exposure", "eyes", "melanoma", "ozone", "radiation", "skin", "ultraviolet", "UV"
428	declining birth rates	"122", "abortion", "AFDC", "birth", "cap", "declining", "Jersey", "month", "rate", "rector"
429	Legionnaires' disease	"Centralia", "disease", "Everest", "hypertension", "IWW", "legionnaire", "patient", "pneumonia", "pulmonary", "wobble"
430	killer bee attacks	"African", "attack", "bee", "garlic", "honey", "immune", "insect", "killer", "nest", "pollen", "wasp"
431	robotic technology	"automatic", "control", "intelligent", "laparoscope", "NASA", "robotic", "surgeon", "surgical", "technology", "telerobot"
432	profiling motorists police	"driver", "interest", "limit", "motorist", "msp", "police", "profiling", "search", "speed", "traffic"
433	Greek philosophy stoicism	"ethic", "Greek", "Jews", "Jewish", "Judaism", "philosophy", "Plato", "Socrates", "soul", "stoicism"

434	Estonia economy	"Baltic", "country", "economy", "Estonia", "Estonian", "GG", "GHG", "Helsinki", "sink", "Tallinn"
435	curbing population growth	"commission", "contend", "curbing", "debate", "Fahrenkopf", "gamble", "growth", "population", "Volstead", "Weyrich"
436	railway accidents	"accidents", "brake", "bridge", "car", "hospital", "injury", "passenger", "rail", "railway", "yen"
437	deregulation gas electric	"competition", "custom", "deregulation", "electric", "energy", "gas", "industry", "power", "restructure", "utile"
438	tourism increase	"attraction", "Australia", "Australian", "increase", "industry", "region", "rural", "tourism", "tourist", "visitor"
439	inventions scientific discoveries	"discoveries", "Edison", "inventions", "inventor", "Nobel", "patent", "phase", "scientific", "STTR", "think"
440	child labor	"BLLF", "carpet", "child", "factories", "Iqbal", "labor", "Nike", "Pakistan", "rugmark", "sweatshop"
441	Lyme disease	"antibiotic", "disease", "disorder", "Gerson", "immune", "infection", "Lyme", "symptom", "tick", "vaccinate"
442	heroic acts	"acts", "adverse", "chamberlain", "heroic", "hood", "injustice", "live", "misfortune", "society", "victim"
443	U.S. investment Africa	"Africa", "African", "apartheid", "foreign", "Gauteng", "gold", "investment", "mine", "rand", "south"
444	supercritical fluids	"chromatography", "dioxide", "extraction", "fluid", "ire", "pressure", "SFE", "solvent", "sulfur", "supercritical"
445	women clergy	"abortion", "Anglican", "Christian", "church", "clergy", "evangel", "marcher", "missionary", "Taleban", "women"
446	tourists violence	"consult", "consular", "Egypt", "Egyptian", "embassy", "Kashmir", "passport", "tourist", "travel", "violence"
447	Stirling engine	"burner", "Cluca", "combust", "cycle", "engine", "heat", "hone", "refrigerate", "Stirling", "thermal"
448	ship losses	"charter", "damage", "liability", "losses", "maritime", "negligent" "recovery", "ship", "shipowners", "vessel"
449	antibiotics ineffectiveness	"antibacterial", "antibiotics", "bacteria", "germ", "Hib", "ineffectiveness", "infection", "microbial", "resistance"
450	King Hussein peace	"Hussein", "Iraq", ""Iraqi", "Israel", "Jordan", "king", "minister", "Palestinians", "peace", "Rabin"

APPENDIX E

CM PARAMETERS

TABLE E.1: CM RESULTS AT M=3, 5, 10, 20

No.	*M*	*d*	Average Precision	Δ%	Precision at 20	Relevant Documents
1	3	10	0.3681	23.23%	0.373	1850
2	3	30	0.3902	30.63%	0.39	1851
3	3	50	0.3968	32.84%	0.403	1839
4	3	100	0.4039	35.22%	0.408	1836
5	3	250	0.4029	34.88%	0.413	1838
6	3	1000	0.3901	30.60%	0.41	1854
7	3	65535	0.3447	15.40%	0.376	1850
8	5	10	0.3731	24.91%	0.388	1850
9	5	30	0.395	32.24%	0.403	1863
10	5	50	0.4037	35.15%	0.408	1857
11	5	100	0.4137	38.50%	0.42	1864
12	5	250	0.4125	38.10%	0.432	1859
13	5	1000	0.396	32.57%	0.421	1865
14	5	65535	0.3554	18.98%	0.376	1840
15	10	10	0.3726	24.74%	0.387	1843
16	10	30	0.3953	32.34%	0.397	1856
17	10	50	0.4035	35.09%	0.411	1854
18	10	100	0.4135	38.43%	0.417	1864
19	10	250	0.4142	38.67%	0.417	1864
20	10	1000	0.402	34.58%	0.411	1872
21	10	65535	0.3543	18.61%	0.386	1867
22	20	10	0.3704	24.00%	0.382	1845
23	20	30	0.3964	32.71%	0.396	1852
24	20	50	0.4006	34.11%	0.406	1859
25	20	100	0.4092	36.99%	0.41	1869
26	20	250	0.4097	37.16%	0.417	1861
27	20	1000	0.3916	31.10%	0.401	1871
28	20	65535	0.3536	18.38%	0.376	1854

REFERENCES

[1] Aljlayl, M. & Frieder, O., 2001. Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation. *ACM 10th Conference on Information and Knowledge Management*, p.295-302.

[2] Alpert, J. & Hajaj, N, 2008. Official Google Blog: We knew the web was big. [Online]. Available at: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [Accessed September 2008].

[3] Baeza-Yates, R. & Ribeiro-Neto. B., 1999. Modern Information Retrieval. *ACM Press, Addison-Wesley*.

[4] Ballesteros, L. & Croft W.B., 1997. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, p.84-91.

[5] Bates, M.J., 2002. After the Dot-Bomb: Getting Web Information Retrieval Right this Time. *First Monday Journal*, 7 (7).

[6] Brin, S. & Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30 (1-7).

[7] Broughton, V., 2005. The Need for a Faceted Classification as the Basis of all Methods of Information Retrieval. *Aslib Proceedings: New Information Perspectives*, 58 (1/2), p.49-72.

[8] Brown, P., Pietra, S.A.D., Pietra, V.J.D. & Mercer, R.L., 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19 (2), p.263-311.

[9] Chau, M., Zeng, D. & Chen, H., 2001. Personalized Spiders for Web Search and Analysis. *Proceedings of the 1st ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, p.79-87.

[10] Chen, J. & Nie, J., 2000. Automatic Construction of Parallel English-Chinese Corpus for Cross-Language Information Retrieval. *Proceedings of ANLP, Seattle*, p.21-28.

[11] Dang, V.B. & Ho, B., 2007. Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining. *Innovation and Vision for the Future, 2007 IEEE International Conference*, p.261-266

[12] Davis, M.W. & Dunning, T.E., 1995. Query Translation Using Evolutionary Programming for Multilingual Information Retrieval. *Proceedings of the 4th Annual Conference on Evolutionary Programming*.

[13] Denton, W., 2003. How to Make a Faceted Classification and Put it on the Web. Miskatonic University Press.

[14] Egypt State Information Service. [Online]. Available at: http://www.sis.gov.eg/ [Accessed 26 February 2008].

[15] EGYPT Toolkit 1.0. [Software]. *Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU)*. Available at: http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/ [Accessed 26 February 2008].

[16] Eiron, N., McCurley K.S. & Tomlin, J.A., 2004. Ranking the Web Frontier. *Proceedings of the 13th International World Wide Web Conference*, p.309-318.

[17] Fellbaum, C., 1990. English Verbs as a Semantic Net. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.278-301.

[18] GNU Wget 1.10.2. [Software]. Available at: http://ftp.gnu.org/gnu/wget/ [Accessed 26 February 2008].

[19] Google, 2002. *Internet Statistics: Distribution of languages on the Internet*. [Online]. Available at: http://www.netz-tipp.de/languages.html [Accessed 26 February 2008].

[20] Gross, D. & Miller, K.J., 1990. Adjectives in WordNet. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.265-277.

[21] Henzinger, M., 2000. Link Analysis in Web Information Retrieval. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, p.45-50.

[22] HTML Text Extractor 1.5. [Software]. Available at: http://www.iconico.com/HTMLExtractor [Accessed 26 February 2008].

[23] Jansen, B.J., Spink, A. & Saracevic, T., 2000. Real life, real users and real needs: a study and analysis of user queries on the web. *Journal of Information Processing and Management*. 36 (2000), p.207-227.

[24] Kleinberg, J.M., 1999. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46 (5), p.604-632.

[25] Kosala, R. & Blockeel, H., 2000. Web Mining Research: A Survey. *SIGKDD Explorations*, 2 (1), p.1-15.

[26] Kraaij, W., Nie, J. & Simard, M., 2003. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*, 29 (3), p.381-419.

[27] Landauer, T.K., 1994. Computerized Cross-Language Document Retrieval Using Latent Semantic Indexing. *Bell Communications Research, Inc.*

[28] Linguistic Data Consortium. [Online]. Available at: http://www.ldc.upenn.edu/ [Accessed 26 February 2008].

[29] Lucarella, D. & Morara, R., 1991. FIRST: Fuzzy Information Retrieval SysTem. *Journal of Information Science*, 17 (2), p.81-91.

[30] Ma, Z., Pant, G. & Sheng, O.R.L., 2007. Interest-Based Personalized Search. *ACM Transactions on Information Systems*, 25 (1), Article 5.

[31] Macdonald, C. & He, B., 2008. *Researching and Building IR applications using Terrier. Proceedings of the 30th European Conference on Information Retrieval (ECIR'08)*.

[32] Miller, G.A. et al, 1990. Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.235-244.

[33] Miller, G.A., 1990. Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.245-264.

[34] Namjoshi, N., 2004. Web Information Retrieval Using Web Document Structures. *M. Sc. Thesis, Graduate Faculty of North Carolina State University.*

[35] Niksic, H. et al, April 2005. GNU Wget 1.10. *Free Software Foundation*. Available at: http://www.gnu.org/software/wget/manual/wget.pdf [Accessed 26 February 2008].

[36] OCLC Online Computer Library Center, 2003. *Summaries: Dewey Decimal Classification*. [Online]. Available at: http://www.oclc.org/dewey [Accessed 26 February 2008].

[37] Ounis, I. et al, 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval.*

[38] Porter, M., 2006. Porter Stemming Algorithm. [Online]. Available at http://tartarus.org/martin/PorterStemmer/ [accessed at 29 February 2008].

[39] Radwan, K., and Fluhr, C., 1993. Fulltext databases as lexical semantic knowledge for multilingual interrogation and machine translation. *Conference of EWAIC'93*

[40] Resnik, P. & Smith, N.A., 2003. The Web as Parallel Corpus. *Computational Linguistics*, 29 (3), p.349-380.

[41] Robertson, S.E. & Spärck-Jones, K., 1976. Relevance Weighting of Search Terms. *Journal of American Society for Information Science*, 27 (3), p.129-146.

[42] Sakre, M., Kouta, M. & Allam, A., 2009. Weighting Query Terms using WordNet Ontology. *International Journal of Computer Science and Network Security (IJCSNS)*. Korea, Seoul. 9 (4), p.349-358. Accessible at: http://paper.ijcsns.org/07_book/html/200904/200904047.html

[43] Sakre, M., Kouta, M. & Allam, A., 2009. Automated Construction of Arabic-English Parallel Corpus. *International Journal of Computer Science and Network Security (IJCSNS)*. Korea, Seoul. (Accepted for publication).

[44] Shahabi, C. & Chen, Y., 2003. Web Information Personalization: Challenges and Approaches. *3rd Workshop on Databases in Networked Information Systems*.

[45] Shen, X., Tan, B. & Zhai, C., 2005. Implicit User Modeling for Personalized Search. *Proceedings of CIKM*, p.824-831.

[46] Sigurbjörnsson, B., Kamps, J. & Rijke, M., 2005. Blueprint of a Cross-Lingual Web Retrieval Collection. *Journal of Digital Information Management*, 3 (4).

[47] Spärck-Jones, K., 1972, 2004. A Statistical Interpretation of Term Specificity and its Application to Retrieval. *Journal of Documentation*, 60 (5), p.493-502.

[48] Talvensaari, T. et al, 2007. Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval. *ACM Transactions on Information Systems*, 25 (1), Article 4.

[49] Terrier 2.1. [Software]. Available at: http://ir.dcs.gla.ac.uk/terrier [Accessed 15 October 2008].

[50] TREC-8 ad hoc and small web topics. [Electronic file]. Available at: http://trec.nist.gov/data/topics_eng/index.html [Accessed 15 October 2008].

[51] TREC-8 small web query relevance judgments. [Electronic file]. Available at: http://trec.nist.gov/data/qrels_eng/index.html [Accessed 15 October 2008].

[52] TREC-8 web collection WT2g. [Electronic web collection]. Available at: http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html [Accessed 15 October 2008].

[53] UDC Consortium, 2003. *Master Reference File Manual*. [Online]. Available at: http://www.udcc.org/mrfmanual.pdf [Accessed 26 February 2008].

[54] Uddin, M.N., Mezbah-ul-Islam, M. & Haque, K.M.G., 2006. Information Description and Discovery Method Using Classification Structures in Web. *Malaysian Journal of Library and Information Science*, 11 (2), p.1-20.

[55] Verma, B. & Zakos, J., 2005. Concept-based Term Weighting for Web Information Retrieval. *Proceedings of the 6th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05), IEEE Computer Society*.

[56] WordNet 2.1. [Software]. Available at: http://wordnet.princeton.edu/obtain [Accessed 26 February 2008].

[57] WordNet Documentation [Online]. Available at: http://wordnet.princeton.edu/man2.1/wnstats.7WN [Accessed 26 February 2008].

[58] Xiaoyi, M. & Liberman, M. Y., 1999. BITS: A Method for Bilingual Text Search over the Web. *Proceedings of Machine Translation Summit VII*, p. 538-542

[59] Youssef, M., 2001. Cross Language Information Retrieval. *Universal Usability in Practice, Department of Computer Science, University of Maryland.*

[60] Zakos, J. & Verma, B., 2006. A Novel Context-based Technique for Web Information Retrieval. *Proceedings of the 9th International World Wide Web Conference*, p.485-503.