Indexing and Weighting of Multilingual and Mixed Documents

Mohammed Mustafa
Digital Libraries Laboratory
Department of Computer Science
University of Cape Town
+27-21 650 4057
mohammed.mustafa@acm.org

Izzedin Osman
Database Laboratory
Department of Computer Science
Sudan University of Science and Technology
izzeldin@acm.org

Hussein Suleman
Digital Libraries Laboratory
Department of Computer Science
University of Cape Town
+27-21 650 5106
hussein@cs.uct.ac.za

ABSTRACT

Non-English-speaking users, such as Arabic speakers, are not always able to express terminology in their native languages, especially in scientific domains. Such difficulty forces many Arabic authors and scholars to use English terms in order to explain precise concepts, particularly when they address technical topics, resulting in mixed/multilingual queries with both English and Arabic terms. Cross Language Information Retrieval (CLIR) allows users to search documents that are written in a language different from the query. However, current algorithms are optimized for monolingual queries, even if they are translated. This paper attempts to address the problem of multilingual querying in CLIR. New techniques that are better suited to the unique characteristics of this problem, in terms of indexing and weighting, are proposed. A new multilingual and mixed test collection containing mixed-language (Arabic and English) computer science documents and mixed-language queries has been created. Experimentally, results show that current CLIR techniques were not designed for these types of multilingual queries and documents and are found to perform poorly whereas the proposed techniques are found to be promising.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval- *Retrieval models*

General Terms

Measurement, Performance, Design, Experimentation

Keywords

Multilingual query, Mixed document, Indexing, Weighting

1. INTRODUCTION

As more users who speak different languages begin participating in the information age, Web content in different languages increases. It is becoming more common to find pages that are available in multiple languages or a single page in more than one language, particularly when documents address technical topics. This is because English content on the Web is being challenged by other languages - Arabic and Chinese are examples. Such non-English languages are growing at a faster rate but at the same time their users show an increasing need for better support for searching the Web [3]. However, despite these growing needs of non-English users, most existing search engines, indexing methods, theories and Web searching techniques are optimised for English and its peer European languages. This is because English remains the primary language on the Web [7]. The majority of credible content on the WWW is available in English. Thus, the support for Web searching for many written languages, particularly from developing countries, is comparatively poor and much weaker than for English. One such difficulty in Web searching for non-English users is the issue of using mixed terms in searching (multilingual querying). A multilingual query is a search query that is mixed between two languages, e.g. the query ' مفهوم الـ Mutual Exclusion' (meaning: concept of Mutual Exclusion) is an Arabic-English multilingual/mixed search query. In a culture where natives use more than one language, especially in scientific domains and their daily business lives, the use of mixed/multilingual terms is very common. Thus, for searching the Web, such natives use mixed languages in order to approximate their information need more accurately rather than using their native-tongue languages in searching.

Current search engines and traditional IR systems perform poorly when handling multilingual querying because, in most cases, they fail to provide the most relevant documents. This is due to two reasons. First, the underlying assumption in IR is that users post queries in their native tongues. Second, most traditional IR systems depend primarily on similarity ranking methods that are based solely on term frequency (TF), document frequency (DF) and inverse document frequency (IDF) statistics, without taking into account the multilingual text in multilingual queries. Ignorance of this feature causes the most dominant documents in the ranked retrieval list to be those documents that contain exactly the same terms as in the multilingual query, regardless of its languages. Figure 1 shows an example of a multilingual query 'ماذا نعني بال Asymmetric key' (meaning: what is meant by Asymmetric key), submitted to the Google Web search engine[1]. Investigation of the retrieved list showed that many monolingual relevant documents, which are written in English, are retrieved at lower ranks while the top ranked documents, which are assumed to be the best, are relatively poor and all of them are multilingual.

SAICSIT '11, October 3–5, 2011, Cape Town, South Africa.

[1] http://www.google.com.

This paper attempts to address the problem of multilingual querying and argues the reasons behind this phenomenon. It shows that current CLIR techniques were not designed for these types of multilingual queries and documents. The paper proposes new techniques to index and re-weight mixed documents so as to make them comparable to monolingual ones and thus result in producing the most relevant documents regardless of the dominant language in the multilingual query or documents. The idea is straightforward and it consists of two steps. Firstly, it combines current architectures of indexing, taking advantages of the benefits of them, and trying to minimize their drawbacks. Secondly, the proposed technique uses a variant of the structured query model [14] so as to re-weight multilingual documents. The approach has many advantages, from the point of view of interpretation, language-awareness and scalability; furthermore it yields high performance for the aims considered in this paper.

Since the first phase of investigating trends of multilinguality in non-English languages, which is not common in current test-beds, is to gather a large corpus for experimentation, a new multilingual and mixed test collection containing mixed-language (Arabic and English) computer science documents and mixed-language queries has therefore been created.

In the following section related work is reviewed briefly. Section 3 describes multilingual querying. In section 4 we discuss in some detail the drawbacks of using current indexing and weighting in multilingual querying. Section 5 describes the proposed methods. Section 6 is dedicated to the experimental setup such as the corpus collection. Section 7 reports on experiments with the proposed methods using the created corpus. Finally the future work and conclusion is provided in the last section.

Figure1. Shows an example of a multilingual query.

2. LITERATURE REVIEW

The issue of using bilingual queries and documents has been discussed in the library community. Hansen et al. [3] enumerated some user requirements for CLIR systems, including the support of multilingual queries and the ability to search multiple languages simultaneously. Rieh and Rieh [10], in their study of Web searching behaviour, concluded that the querying and searching behaviour is dependent on users' needs, purposes of searching and users' ability to speak a foreign language. Thus some users may post queries in their native languages while others prefer to enter multilingual queries. Findings of Lu et al. [6], which were extracted from the analysis of a query log of a search engine and more than 77,000 multilingual queries, showed that mixed query searching between Chinese and English was primarily caused by the followings: using computer technologies, names of magazines and firms; and the fact that some Chinese words do not have a popular translation.

CLIR allows users to search documents that are written in a language different from the query, but neither mixed-language queries nor searches for mixed-language documents have yet been adequately studied. Examples include weighting schemes, indexing methods and ranking functions. In particular, current indexing strategies and weighing methods are designed for indexing several monolingual documents, rather than mixed documents with two or more languages. In CLIR, if the document collection is in more than one language (multilingual collection), then the task is that of Multilingual Information Retrieval (MLIR) [1]. Two major architectures for indexing multilingual collections are centralized and distributed [1] [4]. The centralized architecture considers putting all documents, regardless of their languages, into a single index [8]. Queries are translated into all the target (documents') languages and concatenated to form a single big query, which is submitted to the single mixed collection.

A distributed architecture indexes documents in each language separately [4]. Next, the individual ranked lists are merged into a single ranked list. Different merging methods were proposed. One straightforward merging method is round-robin merging, which interleaves all individual results based on their ranks [22]. Another approach is raw-score merging [22], which sorts all individual results by their original similarity scores. A third approach for merging is the normalized-score merging, which adjusts/normalizes scores before merging [24]. Such normalization can be implemented by dividing each score in each individual list by the maximum obtained score in that list, namely (S_i / S_{max}). A variant of this is to divide each score by the difference between the maximum and the minimum scores obtained in each sub-collection, after subtracting the minimum score in the same list, namely ($S_i - S_{min}/S_{max}-S_{min}$) [24]. The fourth approach is the weighted score merging approach that is based on both document scores and collection scores [12]. The collection scores are based on the CORI (Collection Retrieval Inference Network) [12].

Another type of distributed architecture employed putting all documents into a single unified index, as in the centralized approach [1]. Queries are translated into the documents' languages. Next, a monolingual retrieval, based on the translated queries, is carried out against the unified document index and the individual ranked lists are merged together. In this approach documents in individual lists may overlap due to the use of a single index of documents. The approach in IR studies with such overlapped documents is to sum up the scores of these documents [1]. However, there is an explicit assumption in multilingual information retrieval that documents in individual lists do not overlap.

Regardless of the architecture used for indexing, most CLIR systems use a translation technique in order to perform matching between queries and documents. However, one of the key difficulties in such a process is what to do when more than one translation is known for a given term in a query. In such case a translation disambiguation method is needed. One of the widely-used approaches to this problem is Pirkola's structured query model [14]. The key idea behind the structured query model is that all translations (T_{ij}), in a target language, of a given term (qi) in a source language can be treated as synonyms in the same language of the translations. For example, assume that five Arabic translations are known for the English term 'Object'; then all these five translations are considered as synonyms in Arabic documents. Thus the overall term frequency (TF), which is the number of occurrences of a query term in a document, is computed as the sum of all of the term frequencies for all translations. Similarly, the corresponding

document frequencies (DF), which is the number of documents in which a term appears, for all translations (T_{ij}) of a term (q_i) are combined. Assume that we have k translations for a query term q_i, then TF and DF are computed as follows:

$$TF_j(q_i) = \sum_{i=1}^{k} TF\ (T_{ij}) \longrightarrow (1)$$

$$DF(q_i) = \cup\ DF\ (T_{ij}) \longrightarrow (2)$$

Thus all translations of a given query term in a document are synonyms but in a single target language. One drawback of this approach is that since all translations are treated equally likely, documents containing the most widely-used translations will not be ranked on the top of the retrieved results. Therefore, Darwish and Oard [15] proposed a method, known as Probabilistic Structured Query (PSQ), to incorporate a translation probability in the TF and DF computations - see equations number 3 and number 4. Translation probabilities are usually obtained by making use of different translation resources. For a single resource with (n) translations of a query term q_i, each alternative translation will be assigned a probability of (1/n). Thus, using different resources, the resulting combined probability for a given term q_i would be obtained by summing up all translation probabilities, divided by the number of used resources for translations. The approach causes documents that contain the most likely translation to be retrieved higher than a document with a non-common translation.

$$TF_j(q_i) = \sum_{i=1}^{k} TF\ (T_{ij}) * pr(T_{ij}|q_i) \longrightarrow (3)$$

$$DF(q_i) = \sum_{i=1}^{k} DF(T_{ij}) * pr(T_{ij}|q_i) \longrightarrow (4)$$

where $pr(T_{ij}|q_i)$ is the probability of a translation given a term query q_i. The union operator in equation 4 has been replaced with the sum in order to reduce complexity [16]. However, all of these techniques were originally designed for monolingual retrieval. Since CLIR is a translation process followed by a monolingual retrieval, the models of structured queries are used to perform monolingual retrieval.

Usually such approaches of translation and retrieval are tested by using a standard test collection with an (Information Retrieval) IR system. Several corpora have been developed to serve as standard test collections. However, current corpora can be classified in terms of: single language vs. multilingual; and general vs. specialized [17]. For a given corpus, an overlap in this classification may occur, e.g., a given corpus may be monolingual and specialized.

In terms of their languages, current corpora can be categorized into two types: single language corpora and multilingual corpora [5]. In the single language corpora, all documents are written in a single language. In the second approach, which is multilingual, documents are usually written in several monolingual languages. Multilingual corpora are the most dominant in the standard collections. The most widely known is the different editions of the Text Retrieval Conference (TREC) collections[2]. It contains several monolingual corpora in different languages along with their queries and relevance judgments. Arabic has been included in TREC in 2001 in the crosslingual track. However, most of these corpora are built to help with retrieval of documents based on monolingual queries, even if they are translated. Therefore, most documents are monolingual in several languages. However, some documents are multilingual, e.g. some Japanese documents. Most such multilingual documents present the English translations for some words but not in a tightly-integrated manner.

In terms of vocabulary types, corpora can be classified as general corpora or specialized corpora. A general corpus, as the name indicates, usually contains different genres and domains, such as regional and national newspapers, legal documents, encyclopedias and periodicals. The European Cross Language Evaluation Forum (CLEF)[3], TREC and NII Test Collection for IR Systems (NTCIR)[4] can be considered to be general corpora because test documents in them are general domain news stories [19]. In contrast, a specialized corpus contains terminology in a specific domain. Examples of specialized corpora include NTCIR, which contains some specialized documents, such as in NTCIR-1 and NTCIR-2, which contain abstracts of the Academic Conference papers. More than half are English-Japanese paired documents because abstracts are usually written in English and in Japanese but as parallel text, which is a text in a given language provided with its equivalent in another language.

Most of these specialized documents are either in a single language or constructed from abstracts, not mixed and complete documents with different lengths In addition, Arabic is rare among specialized corpora.

3. WHY MULTILINGUAL QUERYING

Most languages used in developing countries, including the Arabic world, suffer from a limited modern vocabulary. The phenomenon of limited vocabulary and multilinguality has three major reasons. First, is the dominance of English in the scientific domain [7]. Second, many non-English-speaking users, such as Arabic speakers, do not know the exact translations/meanings for most terminology in scientific fields in their native languages. This is because most scientific terminology is borrowed from English and it is not always possible to provide precise translations for new terms, like in medicine and technology. Third, translation/transliteration of newly added terms to a non-English language, such as Arabic, is not usually performed on a regular basis. In addition, scientists who perform the process do not usually invite the experts and scientists in a given scientific domain to participate ("[The Academy of Arabic Language, Sudan Office, personal communication]"). As a result, Arabic words are ambiguous, chaotic and are almost not understood by Arabic speakers.

Though the English part of the multilingual query may have a proper translation in Arabic, science scholars sometimes do not prefer to use such a translation in their communications or for searching across documents. This is because of the regional variation difficulty, especially in scientific terminology. Unlike in the news genre, the problem of regional variation in scientific domains is crucial, especially when considering regions like the Middle East or the Arabic-speaking world. The latter region has more than 21 countries. As a result, scientific modern terms in Gulf countries may be totally different from those in Levantine countries. Such problems forced many Arabic authors and lecturers to use English terms in order to explain precise concepts. On the Web, the problems result in a trend of using multilingual querying and multilingual documents in both English and the native languages. A multilingual document is written in two languages.

[2] http://trec.nist.gov/

[3] http://www.clef-campaign.org/

[4] http://research.nii.ac.jp/ntcir/index-en.htm

This natural human tendency is very common in the non-English-speaking world. It is caused by the fact that many people are able to express some keywords in languages other than their native tongue, e.g., scientific English terms vs. Arabic for Arabic speakers. The typical Arabic speaker speaks a mixture of tightly-integrated words in both English and Arabic (and various slang variants) that will muddle most algorithms in IR. Students at Arabic universities may ask a question like 'Deadlock ما هوالـ ', which is a tightly-integrated question that is presented in two languages and means 'what is deadlock' instead of 'ما هو الإستعصاء' because terms like deadlock are more meaningful and unambiguous to them. Examples include lectures where some text is best expressed in an indigenous/home/local language while other text may best be expressed in a variant of English. For such non-English users, multilingual querying may be more appropriate because this is often the best and the only balanced way to fill the gap between the limited vocabulary and searching needs.

Most weighting algorithms, indexing methods and ranking approaches of current search engines and traditional IR systems are optimized for monolingual queries, even if they are translated, and documents and were not designed for such multilingualism in queries and documents. This underlying assumption causes the most dominant documents on the ranked retrieval list to be those documents that contain exactly the same terms as in the multilingual query, regardless of its languages. Thus, weighting of terms in the Arabic portion of multilingual queries is handled in a similar way to English term weighting. Given these trends and the need for relevant information by users in developing countries, it is essential to develop algorithms for future search engines that will allow non-English-speaking users to retrieve relevant information created by other multilingual users.

4. WHY CURRENT INDEXING AND WEIGHTING ARE NOT OPTIMAL

4.1 Indexing Techniques

It is illustrated above that there are two architectures for indexing. However, neither of them is the optimal solution for indexing multilingual/mixed documents. Firstly, centralized architecture appears adequate for indexing multilingual documents, because of making use of a single index, but it has been shown to have some problems [5] .One major problem with centralized architecture is that index weights are usually overweighted. This is because the number of documents (DF) increases while the number of occurrences of a term (TF) is kept unchanged and thus weights are overweighted. For example, consider a collection containing 6,000 monolingual Arabic documents along with 70,000 documents in English. In a centralized architecture the (N) value (number of all documents) in the (IDF) of a term, which is computed as *log(N/DF)* in order to estimate a term importance, for the Arabic collection will increase to 76,000, instead of 6,000 when all documented are placed together in a single collection. This will cause weights of terms to overweight and thus documents with small collections are preferred. Turning our attention to multilingual querying, it is apparent that the same phenomenon of overweighting would occur in its worst case if a centralized index is used. The work of Lin and Chen [5] did not consider the number of occurrences of a term in different languages, as in multilingual documents, although such documents were addressed in the study. This is clear from their conclusion about the major drawback of a centralized index, which is the overweighting. The study stated clearly that only the document frequency increases but the term frequency does not. The same assumption appears in a major proportion of experiments in the literature. Obviously, in multilingual documents both the number of documents and the number of occurrences of a term increase. Moreover, the weight of each term of the generated multilingual queries in the centralized architecture is computed independently, regardless of its language. This is a crucial problem for multilingual queries and documents. For example, consider a query consisting of the word: 'Inheritance'. In the centralized architecture, this query would be 'Inheritance الوراثة', in which the Arabic translation' الوراثة' is concatenated to the original query. In this query the centralized approach computes the weight of each of the two words independently though they are similar words but in different languages.

The second approach to indexing is the distributed approach. With respect to multilingual querying, it is clear that the dominant approach in distributed architecture is to translate a user query to target language(s) and next a monolingual language-specific search is carried out per each sub-collection followed by a merging method. Distributed architectures provide users with two options to handle multilinguality. The first option is to divide – even if implicitly using tools - each multilingual document, according to its languages, across all/some of the language-specific sub-collections. Such an approach probably causes multilingual documents to lose their information richness and meanings, especially if their text, which is written in multiple languages, is tightly-integrated. Thus, when a multilingual query is submitted to a single language sub-collection, multilingual documents would not compete because only a small part of terms in the multilingual query will appear in these partitioned multilingual documents. This is because the number of occurrences of terms in a given language sub-collection is expected to be low for multilingual queries and thus scores between monolingual documents and multilingual documents would not be comparable.

The second option that could be applied for multilingual documents by the distributed approach is to index all documents in a single unified big index, as in the centralized approach. Next, queries are translated into the documents' languages. Thus, a monolingual retrieval is implemented for each query in a target language against the single index. Then, a merging process to obtain the final list is applied. At first this method sounds more adequate for multilingual documents. But such documents, which are almost multilingual, may be ranked on more than one individual list, meaning an overlapping probably would occur. However, the assumption in IR studies, as mentioned above, in such overlapped documents is that: since these documents appeared in more than one list, then they are more likely to be relevant than those appearing on a single list and thus such documents should be ranked higher. One approach to apply such methodology is to sum up the scores of these documents. This will rank most multilingual documents at a higher level and thus decrease the effectiveness of the IR system.

From these trends it is concluded that current indexing approaches are not optimal for multilingual querying and documents

4.2 Weighting Methods

Current CLIR has focused on developing approaches for effective translation of queries and/or modification of weighting methods. This is because CLIR process is substituted by a translation process followed by a monolingual information retrieval. Thus, most current methods were originally designed

for monolingual retrieval. For instance, let's consider the probabilistic structured query (PSQ), which was explained above. This model was primarily developed for an effective translation and retrieval in CLIR. The conclusion of this model, and almost all of its different variants, is based on monolingual retrieval. However, CLIR techniques had proven their ability to retrieve and rank news stories but this does not mean that they are ready to be applied to scientific domains, especially in multilingual non-English collections. In such cases of documents, there is a possibility of poor retrieval because the assumed language model is incorrect. To conclude, it is possible to say that current CLIR weighting methods are Language-un-aware solutions because they just shrink the CLIR process to a monolingual IR by translating the user query.

5. THE PROPOSED INDEXING ARCHITECTURE AND WEIGHTING

The proposed technique is straightforward and its basic idea consists of two steps. Firstly, it combines centralized and distributed architectures, taking advantages of their benefits, and trying to minimize their drawbacks. For the centralized architecture, the proposed technique minimizes the overweighting drawback. This is done by indexing multilingual documents only in a centralized architecture, instead of indexing both monolingual and multilingual documents. We will call such indexing architecture through the rest of this paper the '*centralized-multilingual-sub-collection*'. Moreover, instead of partitioning multilingual documents and/or overlapping them in each individual list, the proposed technique uses the distributed architecture for monolingual documents only, but not for multilingual ones. We will call this indexing architecture the '*distributed-monolingual-sub-collection*'. The basic technique is illustrated in Figure 2.

Secondly, the proposed technique uses a variant of the structured query model to re-weight documents into the centralized-multilingual-sub-collection only. Two reasons exist for using this modified variant of the structured query model. These reasons are as follows: to avoid overweighting again because of the use of the centralized architecture, partially, in the combined index; and to re-weight multilingual documents that contain the same terms but in different languages. This is done by identifying synonyms in multilingual documents across languages rather than synonyms in a single language. The latter is used in the original structured query model and PSQ to avoid exact matching between multilingual query terms and documents' terms because such matching results usually in ranking multilingual documents at the top of the retrieved list; and to make multilingual documents comparable to those monolingual ones that are indexed in the distributed-monolingual-sub-collection(s). Details about the proposed techniques are illustrated below.

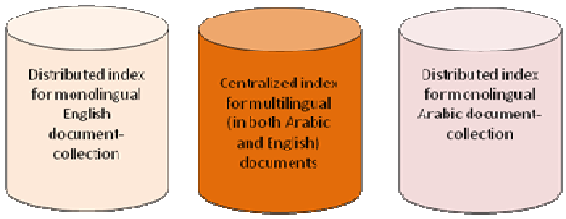

Figure 2. The combined index.

5.1 Combined Indexing

A typical distributed architecture does not prefer collections with small number of documents, as in a centralized architecture. The retrieval performance of each monolingual run is much better than in a centralized architecture. This is because both queries and documents in each distributed sub-collection index are in the same language. Therefore, the proposed combined index creates a distributed-monolingual-sub-collection for each language that is used in monolingual documents only, but not for documents in multiple languages. Thus, multilingual documents are not included in these distributed-monolingual-sub-collection(s). The significant benefit of indexing monolingual documents only in distributed architectures is the efficient retrieval in each sub-collection, due to the similarity in languages between queries and documents. In addition, since multilingual documents are not included in the monolingual indexes, partitioning these documents as well as overlapping of them in individual lists were avoided, unlike the normal distributed index which doesn't consider the multilingualism feature in multilingual documents.

On the other hand, a centralized architecture maintains indexing, searching and retrieval in a single index, regardless the used languages and thus no merging process is required. This feature in centralized architecture is very helpful for multilingual documents because retrieval from a single index is expected to perform better than individual retrieval followed by a merging process. Therefore, the better approach to deal with multilingual documents, but not documents in several monolingual languages, is to from a unique centralized index, regardless of the used languages in these multilingual documents. Indexing multilingual documents only in a centralized sub-collection index has different advantages: First, it avoids partitioning these documents across several distributed indexes and hence, so information richness inside multilingual documents will not be lost; second, it minimizes the overweighting problem in the centralized architecture to its lowest level because monolingual documents are not included in this centralized-multilingual-sub-collection index. Thus, the number of documents in the collection (N) will not increase and consequently the IDF for a query term will be kept as it should be.

Despite the use of the centralized index, the entire architecture of the proposed strategy is distributed. The index can be viewed as a big combined repository for indexing a centralized sub-collection that is inserted on the same level with other distributed (monolingual) sub-collections. One significant advantage of using this proposed indexing architecture is that it can be easily adapted to other languages, not only Arabic and English.

5.2 Weighting in Mixed Sub-collection

Although the overweighting drawback in the centralized-multilingual-sub-collection is mitigated as shown above, documents still may overweight. Therefore, a re-definition for the structured query model is proposed. While the structured query model could help in retrieving documents that contain all translations of a query term by considering them as synonyms in the same language, it cannot be employed for multilingual documents unless it is re-defined, because it is originally developed for monolingual documents. The same assumption holds for probabilistic structured query.

Usually text in scientific documents in non-English languages, Arabic for example, is presented in a strongly and a tightly-integrated manner. In particular, Arabic scientific documents

usually contain different terms/portions/ snippets/phrases/paragraphs in two languages –English is one of them- but in a tightly-integrated text. For instance, consider a document that contains the English term 'inheritance' 7 times and in other positions the same document contains the Arabic translation(s) for the same term ('الوراثة' and/or 'التوريث') 9 times. Now, consider a multilingual query Q = ' مفهوم الـ Inheritance' (meaning: concept of inheritance). This document will not consider the English term 'Inheritance' and its Arabic translations as synonyms across the two languages. This means that the TF for each term(s) in a given language will be computed independently when the query is translated, despite the fact that the TF should be (9+7) regardless of the used language (s) in the document.

To avoid such problematic weighting encountered in multilingual documents, a variant of structured query model is proposed in order to re-weight technical terms in multilingual queries. This is done as follows: if a translation (T_{ij}) for an English technical term (q_i) appears in a multilingual document (D), we treat this as if the query term (q_i) occurs in the same document (D), and hence, both the translation T_{ij} and the term q_i are considered as synonyms but in different languages. With respect to DF, if document (D) includes a translation T_{ij}, we can treat that document as if it contains the query term q_i and vice-versa. Turning our attention to translated Arabic terms and how their weights are computed in multilingual documents, their weights are neutralized. This is very important in order to avoid overweighting again. Assume that we have *k* translations for a query term q_i, then TF and DF are computed as follows:

$$TF(q_i) = \sum_{j=1}^{k} TF\,(T_{ij}) + TF\,(q_i) \longrightarrow (5)$$

$$DF(q_i) = \cup\ DF\,T_{ij} \ \cup DF\,(q_i) \longrightarrow (6)$$

These modified weights make no use of translation probabilities, unlike the probabilistic structured query, and each candidate translation is considered as being equally likely. This is due to the fact that translations of technical jargon are not similar to those in the news domain, which was used to test PSQ. In general-domain, such as news, it may be adequate to retrieve documents that contain the most probable translation, among a set of synonymous translations. In technical topics and documents the criteria does not hold, especially for a language with several regional variations – as in the Arabic world. Consider the Arabic translations for the technical English phrase 'object oriented programming', which are:'البرمجة الشئية ', 'البرمجة كائنية التوجه', 'البرمجة موجهة الأهداف' and ' البرمجة كائنية المنحى'. All these alternative translations can be used in scientific Arabic documents, but according to the dialect of the writer. Hence, occurrence of superfluous translation, e.g.' البرمجة الشئية', in documents does not mean that these documents are irrelevant. Due to this important fact, above equations makes translations equally likely.

Retrieval is straightforward. Multilingual queries are bi-directionally translated. Translation details are explained in the experimental setup section. The monolingual Arabic queries will be used to retrieve Arabic documents from the Arabic distributed-monolingual-sub-collection. The English ranked lists will be obtained by running the translated monolingual English queries against the English distributed-monolingual-sub-collection. For the multilingual-centralized-sub-collection, both translated Arabic and English queries are concatenated, as in the normal centralized architecture, to form a big query. Weights are modified in the index and then an individual result list is obtained. Finally one of the merging methods, which are illustrated above, is used to obtain a final ranked list. Figure 3 illustrates the entire proposed approach.

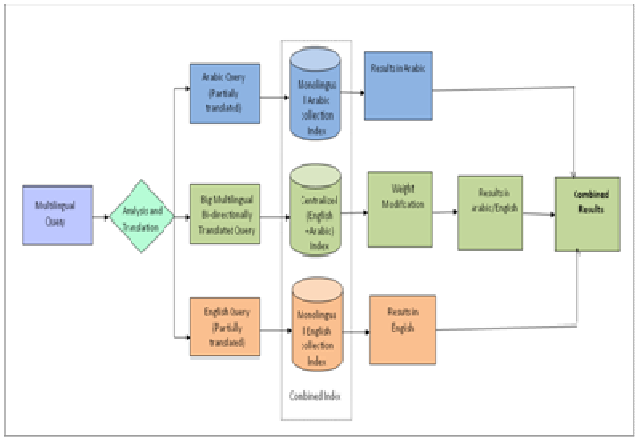

Figure 3. The proposed solution.

6. EXPERIMENTAL SETUP

There are inherent difficulties when conducting experiments presented in this paper. One of the most important difficulties was the lack of a standard test-bed with multilingual and mixed documents, particularly in both Arabic and English languages. The first choice is to use one of the standard test-beds, like TREC or CLEF, but most currently available test collections and almost all CLIR collections have focused upon general-domain news stories [19]. Moreover, the majority of these collections is monolingual or consists of several monolingual documents in different languages, with each document in a given language, rather than documents with different portions/terms that are tightly–integrated in multilingual forms. Thus, developing a multilingual Arabic/English corpus of common computer science vocabulary was the first step among the experimental setup activities.

6.1 Corpus Building and Processing

The corpus has been collected both automatically and manually. In the automatic gathering process, the WebReaper[5] Web crawler was used to create local copies of sites, which contain books, lectures and articles on common computer science. The choice was mainly governed by the availability of computer science documents. A manual collection of data was also considered. The collected documents were merged into a single pool. Duplicates were removed and a total size of 6.1 GB of raw Web-based data was obtained. The process was characterized by a major challenge, which is the issue of respecting copyright and intellectual properties. Thus, an iterated process of contacting books' authors was carried out before collecting documents.

Next, documents were processed. At first, documents in different formats (shtml, html, doc, pdf, etc) were converted to HTML. During this step, tags, symbols, images and special characters, like ®, were removed. Only the raw text was retained. The new formatted HTML documents were saved in a common encoding, which is Unicode. Along with this step, each document was tagged with a special tag for referencing purposes. A normalization process was carried out for all documents. For instance, English words were normalized in terms of case-sensitivity while Arabic words were normalized to control the orthographic variations and spelling variants, which

[5] http://www.webreaper.net/

are very common in Arabic [21]. Thus, different glyphs for some letters are rendered to a single letter. For example, the letters ALIF HAMZA with its two glyphs (أ،إ) and ALIF MADDA (آ) were replaced with a bare ALIF (ا). Regional variants in the collection were kept. In order to prepare the text for multilingual indexing later, every word/phrase/portion/ paragraph - depending on how much a document is mixed - in documents was marked with a language tag attribute using a simple language identifier. This would help to identify the correct stemmer during the indexing phase.

6.2 Corpus Statistics

In order to obtain the essential information needed for the corpus statistics, the *Lucene* IR system was used. Lucene is a high-performance, full-featured text search engine library written entirely in Java[6].

Table 1. Collection summary.

Description	Language(s) in documents	Number of words	Total
Number of words	English words	10,641,909	14,789,137
	Arabic words	4,147,228	
Number of distinct words	Distinct words in English	208,635	281,358
	distinct words in Arabic	72,723	
Number of documents	Monolingual English documents	18,451	23,811
	monolingual Arabic documents	137	
	multilingual (both Arabic and English) documents	5,223	
Average number of words per document		612	

The size of index in Lucene is roughly 20-30% compared to the size of text to be indexed. Thus, using Lucene simple statistics about the numbers of words in the collection were extracted. Table 1 shows these statistics. From the table, it is observed that the average number of words per document is relatively high. This is because the collection is specialized, unlike in news. Another important observation is that although the data has been collected arbitrarily, monolingual Arabic documents are very rare, at least in terms of common computer science. It is also observed from the table that Arabic text has more words occurring only once.

6.3 Topics/Queries Construction and Relevance Judgments

It is known that queries for tests should be representative of the queries submitted by users of the target application [2]. This approach is followed in this experiment. Hence, to generate queries for the experiment, the selected potential users were a group of 25 students at different academic levels at an Arabic-speaking university. Around 125 queries were obtained. All submitted queries were pooled into one set. Duplicates and semi-similar queries were removed. Hence, a cleaned set of 67 queries was obtained. An important note was observed in the submitted queries: more than 68% of these queries, before pooling, were expressed in multilingual forms. A set of 25 multilingual queries was selected. The selection of queries was based on a suitable recall: most queries should have suitable relevant documents. All of the selected queries are multilingual. Table 2 lists some of these queries with their translations/meanings in English. Queries were numbered (DLIB01-DLIB25) for referencing purposes. The average no. of words per query was found to be 4.3 with 1.9 and 2.4 as the average number of words for Arabic and English, respectively.

Table 2. Sample queries.

Query #	Query	Counterpart in English
DLIB06	Deadlock مفهوم الـ	Concept of deadlock
DLIB07	Secure ماذا نعني بالـ Socket Layer	What is meant by Secure Socket Layer
DLIB09	Interpreter الفرق بين ال و ال Assembler	Difference between Interpreter and Assembler
DLIB15	Polymorphism شرح الـ	Explain Polymorphism
DLIB21	Entity مثال في ال Relationship Model	Entity and Relationship Model, Example
DLIB25	Data Mining تقنيات ال	Data Mining Techniques

Queries were put in files that are identical to TREC topics (queries) format. In TREC, each topic contains three fields: title, description and narrative. The title field is supposed to be a short query which is usually typical to those on Web applications. The description field is a longer version of the query. The narrative field describes the criteria for relevance. Thus, in this work the same fields were used plus two extra fields: creator and original Query. The creator filed is added for referencing purposes, i.e. the name of a query's creator, while the originalQuery field stores the original mixed multilingual query. The Arabic versions of query files were translated by some volunteers.

Most TREC conferences are based on binary relevance assessments (*relevant*, *not relevant*) with a very low threshold for accepting a document as relevant. Due to the nature of our experiments, it is obvious that binary relevance cannot reflect the possibility that documents may be relevant to a different degree because some scientific documents may contribute more information to the query while others may contribute less without being totally irrelevant. Thus, multiple levels of relevance (Graded Relevance) were used to assess documents that were retrieved by the queries. Relevance assessment is done on a six-point scale (0-5), with 5= highly relevant document, 4= fairly/suitable relevant document, 3= low relevance document, 2= marginally relevant document, 1 = possibly not relevant document and 0 = irrelevant document. It is well known that the quality of the relevance judgments has a major impact on retrieval process. Therefore, two PhD's and three Master's students in computer science were requested to assess the relevance of documents. The bottom line is that we do not claim to have constructed a standard test collection, but rules in building such collection were used as guidelines. It is important here to mention that the corpus was validated in terms of terms frequencies, sparseness and growth, using statistical tests but this is beyond the scope of current work.

[6] http://www.lucene.com

Since our experiments' task emphasizes top ranked documents, the Discounted Cumulative Gain (DCG) measure was used. DCG is a performance measure metric that is becoming increasingly popular for evaluating Web search engines and related applications [21]. The assumption in this measure is that: the greater the ranked position of a relevant document, the less valuable it is for the user, because it is less likely to be examined by users, or cumulated information from documents already seen. Thus, the DCG uses graded relevance as a measure of the usefulness or gain from examining a document. The computation of the DCG is performed as described by [13]. GCD is usually computed at a given rank (p), which is similar to precision at rank p. Since the focus of the measure is on the top ranks, the values of *p* are typically small, such as 5 and 10 [2]. Thus, in the experiments the assessors assessed the top 10 documents for each query.

6.4 Indexing and Stemming

Indexing of documents in the corpus was done using Lucene. Four fields were proposed: <TITLE-Arabic>, <CONTENTS-Arabic>, <TITLE-English> and <CONTENTS-English>. The <TITLE-Arabic> and <CONTENTS-Arabic> fields were used for the Arabic distributed-monolingual-sub-collection, while the <TITLE-English> and <CONTENTS-English> fields were used for the English distributed-monolingual-sub-collection. This is similar to the use of the <TITLE> and <TEXT> fields in TREC documents. For the multilingual-centralized-sub-collection, all of the four fields were used. This is due to the multilingualism characteristic in mixed documents. Thus, depending on the document' language(s) and the type of indexing in the combined index, some or all fields may be used.

Since documents are often divided into fields, a linear combination of the scores that are obtained from scoring each field is computed. However, using such combination is criticized by Robertson, et al. [9], who proposed an extended version of the BM25 [11] for fields weighting scheme. BM25 is a probabilistic IR model that has widely adopted by the IR community. The extension of the BM25 in Robertson's work is based on refraining of doing linear combination of scores obtained from scoring every field in documents. The idea is to combine the term frequencies of the different fields in order to compute a single score for the entire document and thus the scoring function is applied only once to each document, although the document is structured. Thus, in experiments this extended model of BM25 was applied. During the indexing phase, words in documents were stemmed and stopwords were removed. Light10 stemmer [20] was adopted for Arabic words while snowball stemmer [24] was used for English ones.

6.5 Query Translation

In the experiments, multilingual queries are used as source queries. For the implementation they should be bi-directionally translated. Bi-directional translation here means that English words are translated to Arabic and vice versa. Since English words in multilingual query are assumed to be technical terms in computer science, an English-Arabic computer-based dictionary, collected from different sources, was used. For the Arabic words in queries an application was developed so as to use Google translator. Usually such Arabic words in multilingual querying are general vocabulary words, unlike English words in multilingual queries, which are usually good candidates and useful clues for searching.

7. EXPERIMENTS AND RESULTS

In the first phase of experiments, four runs were submitted. The four runs used the <*OriginalQuery*> field, which is the multilingual query, as source queries. The source multilingual queries are bi-directionally translated into Arabic and English, as explained in the experimental setup section. Both translated queries of each source query are concatenated to form a third big single query. The bi-directional translated versions of queries were kept. The source queries themselves were not used in retrieval. The three queries (monolingual Arabic queries, monolingual English queries and the big concatenated queries) were used to retrieve the corresponding document sub-collections, according to the proposed combined architecture of indexing. Then, modifications of weights as described by section 5.2 are adapted to re-weight documents in the centralized-multilingual-sub-collection index only. Next, different merging methods were implemented. In all runs the top 1000 documents were chosen as a final result. The different merging methods for the four runs were:

1. Cmb01A: in this run, result lists were merged by the raw score merging method.
2. Cmb02A: the CORI merging method was used to merge the intermediate result lists into a single list.
3. Cmb03A: in this run, scores in each individual ranked list were normalized by the difference between the minimum and the maximum scores after subtracting the minimum score of the original score.
4. Cmb04A: in this run, maximum scores were used for normalizing original scores in each individual result list.

In order to compare the effectiveness of the proposed solutions, an experiment that uses the original centralized architecture was also conducted as a baseline. This widely reported baseline seems appropriate in this case because usually most experiments consider putting all documents together in a single index. Thus, another separate index was created to index all documents in the corpus in a single big index, regardless of their languages. Next, the big concatenated queries were used to retrieve documents from this big index file. This is typically what the original centralized index does. Neither the proposed combined indexing nor the proposed modified weighting were used in this run. We called this run as *CntBL*.

Although, this paper is aware of the use of both the combined architecture and the modification of weighting together, it was interesting to show how the latter, meaning weight modification, performs independently. Such an experiment will provide us with an indication of which approach contributes better to the retrieval enhancement: combined index or/and modification of weights. Therefore, in the second phase of experiments, another run that considers putting all documents together, regardless of their languages, in a single index was conducted. In this experiment, modifications of weights as described in section 5.2 were adopted. The proposed combined architecture was not used and thus no merging method is needed. For the retrieval, the big concatenated queries were only used. We called this run as *CntMW*.

Figure 4 shows the average DCG, across the 25 queries, @ top k documents (k = [1, 3, 5, 8, 10]) for each of the six runs. Table 3 reports the same results in tabular formats. Results show that the proposed solutions yield the best retrieval in all runs and the improvement in retrieval is highly significant (p-vale < 0.05). This is due to the use of the combined architecture of indexing, which minimizes the effect of the overweighting and avoids partitioning and/or overlapping of multilingual documents. Further, synonymy across languages in multilingual documents affects both the number of terms (TF) and the number of

documents in which the term occurs (DF). Consequently, most of the top ranked documents are not multilingual, unlike in the baseline run. In particular, instead of considering translations of a query term as synonyms, the baseline method computes each term weight independently and as if it has no synonym in another language. Such approach hurts retrieval effectiveness.

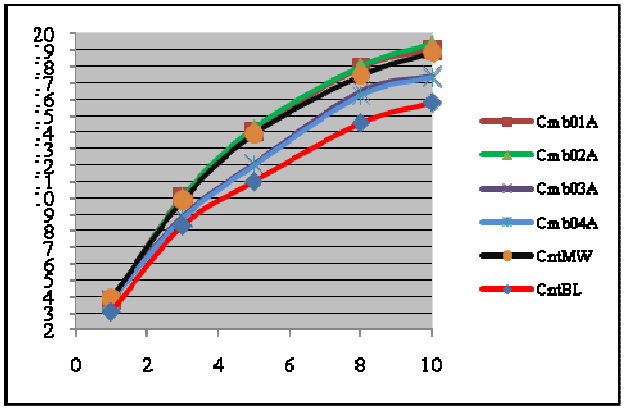

Figure 4. Average DCG @ top [1,3 ,5 , 8, 10] for all runs.

Although experiments emphasize the highly relevant documents task, it was observed that the retrieval performance of the proposed solutions is still much better than the baseline, even after the first 10 documents. Thus, results are also an indicator of the fact that English is still the dominant language of science, at least when it is compared to Arabic.

The best retrieval effectiveness @ top 10 documents is obtained by the proposed solutions engaged with CORI. Although comparing the effectiveness of the used merging methods is beyond the scope of current work, this is probably because CORI normalizes raw scores of individual lists according to the average score of the corresponding sub-collection and, hence, top documents in collections with low number of documents will not be favoured. The performance of the proposed solutions with raw score merging is good. This is probably because the same IR model and the same term weighting scheme were used for all sub-collections.

The proposed solutions with raw-score and CORI methods are almost similar (the difference is statistically insignificant).

Table 3. Average NDC for all runs.

Run	Average DCG @ top				
	1	3	5	8	10
Cmb01A	3.73	10.07	14.02	17.89	19.02
Cmb02A	3.62	10.02	14.21	17.98	19.34
Cmb03A	3.85	8.86	12.05	16.47	17.41
Cmb04A	3.79	8.73	11.95	16.22	17.32
CntMW	3.89	9.86	13.89	17.46	18.85
CntBL	3.06	8.30	10.99	14.57	15.81

They produced 20.3% and 22.33% relative improvement in retrieval performance, respectively, when they are compared to the baseline run. The performance dropped down after considering the proposed solutions as well as normalizing scores in individual lists by the minimum/maximum and maximum corresponding scores (Cmb03A and Cmb04A runs). The reason behind this degradation in performance did not inherit from the proposed solutions. This is probably associated to the merging strategies in these two methods, which have some drawbacks. For example, if the maximum score in a ranked list is much higher than the maximum score in a second list, both the top two scores will be normalized to 1 and all individual scores will be normalized in a range between 0 and 1. Thus, each individual list will be having at least some of its top documents in the final ranked list after merging, although one of the lists has documents with lower scores in the top ranks. The same criteria hold for reasoning about why retrieval of both the two methods obtains the highest values @ top 1 document. However, despite this drawback the retrieval effectiveness of both methods is much better than the baseline when they are integrated with the proposed solutions. The retrieval improvements are statistically significant ($p < 0.05$).

The proposed solutions consist of two approaches, each of which contributes to the performance improvement differently. This is obvious when comparing the retrieval effectiveness between CntMW, which considers only weight modifications with a centralized architecture, and first four runs, which adopted both the modifications of weights across synonyms and the combined architecture. In particular, although the combined architecture has a big role in the final enhancement of retrieval, modifications of weights contribute much more. Therefore, results of CntMW are much closer to those in Cmb01A and Cmb02A. However, differences in retrieval improvements are statistically insignificant.

8. CONCLUSION AND FUTURE WORK

The problem of querying, indexing and weighting of multilingual queries and documents is critical and may affect retrieval performance. The focus of this paper is therefore on development of new techniques that are better suited to the unique characteristics of this problem. The proposed techniques are based on indexing a centralized sub-collection that is inserted (distributed) among other distributed sub-collections. Next, a variant of the structured query model handles synonymy across languages. Since most corpora are built from news, legal documents and encyclopedias, a new multilingual test collection containing mixed-language (Arabic and English) computer science documents and mixed-language queries has been created for testing. Results showed that performance of the proposed solutions with some merging methods is much better than using a centralized index. However, using the appropriate merging method affects retrieval performance greatly. Results also showed that modifications of weights in multilingual/mixed documents contribute to the performance enhancement much more than the use of the proposed combined indexing.

Future work will focus on investigating and extending other weighting schemes to handle multilinguality more accurately. To achieve this goal, multilingualism in scientific Arabic documents will be studied in depth. Merging methods will be also investigated, e.g. a new merging method based on logistic regression. According to our knowledge, till now there is no logistic merging model that incorporates the parameter of how much a document is multilingual, although logistic regression is well studied.

9. ACKNOWLEDGMENTS

Our thanks go to the CLIP research group at the University of Maryland – College Park, USA for their valuable comments. This research is supported by the University of Cape Town, the Telkom/NSN/Telesciences/THRIP Center of Excellence and the National Research Foundation.

10. REFERENCES

[1] A. Chen, F. Gey, "Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding", Journal of Information Retrieval, 2004, 7(1-2), 149-182.

[2] Croft , B., Metzler, D., and Strohman, T. 2009 Search Engines: Information Retrieval in Practice. Addison-Wesley.

[3] Hansen, P., Petrelli, D., Karlgren, J., Beaulieu, M., and Sanderson, M. 2002. User-centered interface design for cross-language information retrieval. Proceedings of the twenty-fifth annual International ACM SIGIR Conference, ACM Press, New York, NY, 383–384.

[4] K. Kishida, "Technical issues of cross-language information retrieval: a review", Journal of Information Processing and Management, 2005, 41(3), 433 – 455.

[5] W. Lin, and H. Chen, "Merging Mechanisms in Multilingual Information Retrieval", Advances in Cross-Language Information Retrieval LNCS, 2003, 2785, 175-186.

[6] Lu, Y., Chau, M., Fang, X., and Yang, C. C. 2006. Analysis of the Bilingual Queries in a Chinese Web Search Engine. Proceedings of the Fifth Workshop on E-Business (2006, Milwaukee, Wisconsin, USA).

[7] Miniwatts Marketing Group (2011), "Internet World Stats Usage and Population Statistics", Available at: http:// www.internetworldstats.com/, Last accessed 20 -4- 2011.

[8] J. Y. Nie, and F. Jin, "A multilingual approach to multilingual retrieval". Advances in cross-language information retrieval LNCS, 2003. 2785, 101–110.

[9] Robertson, R., Zaragoza, H. and Taylor, M. 2004. Simple BM25 Extension to Multiple Weighted Fields. Proceedings of CIKM Conference (Washington, DC, USA, November 8-13) CIKM'04, ACM Press, New York.

[10] H. Rieh, S. Rieh, "Web Search across Languages: Preference and Behavior of Bilingual Academic Users in Korea", Journal of Library & Information Science Research, 2005, 27(3), 249-263.

[11] Robertson, S. E. Walker, S. Some simple effective approximations to the 2-Poisson model for probabilistic Weighed retrieval. 1994. In Proceedings of the 17th Annual International SIGIR Conference, Springer-Verlag, 245-354.

[12] Callan, J.P., Lu, Z. and Croft, W. B. 1995. Searching distributed Collections with inference network. In Proceedings of the 18th Annual International ACM SIGIR Conference (Seattle, WA, USA) ACM Press, 21-28.

[13] Järvelin, K., & Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Information Systems, 20(4), 442− 446.

[14] Pirkola, Ari. 2003. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference, ACM Press, 55-63.

[15] Darwish, Kareem and Oard, Douglas. 2003. Probabilistic structured query methods. In Proceedings of the 21st Annual International ACM SIGIR Conference, ACM Press, 338-344.

[16] K. L. Kwok. 2000. Exploiting a chinese-english bilingual wordlist for english-chinese cross language information retrieval. In Proceedings of the 5th International Workshop on Information Retrieval with Asian languages (2000), 173–179,

[17] McEnery, T., Xiao, R., and Tono, Y. 2006 Corpus-based language studies: an advanced resource book. Routledge.

[18] F. Gey, C., K. Noriko, C. Peters, "Language Information Retrieval: the way ahead", Journal of Information Processing and Management, 41(3), 415 – 431.

[19] Rogati, M. and Yang, Y. 2004. Resource Selection for Domain Specific Cross-Lingual IR. In Proceedings of ACM SIGIR Conference SIGIR'04, ACM Press, NY.

[20] Larkey, S. L., Ballesteros, L., and Connell, E. M. (2005), Light stemming for Arabic information retrieval. Arabic Computational Morphology: Knowledge-based and Empirical Methods.M. Tayli, and A. I. Al-Salamah, "Building microcomputer systems", Communications of the ACM, 1990, 33(5), 495-505.

[21] Voorhees, E. 2001. Evaluation by highly relevant documents. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.). Proceedings of the 24th annual international ACM SIGIR conference ACM Press, New York, NY, 74-82.

[22] Kraaij D. Hiemstra, R. W. Pohlmann, and T. Westerveld, 2001. Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In C. Peters (Ed.). Cross-language information retrieval and evaluation. Lectures in computer science Springer Verlag 2069, Germany, 102-115.

[23] Powell, A., French J., Callan J, Connell, M. and Viles, C. 2000. The impact of database selection on distributed searching. In Proceeding of the 23rd Annual International ACM SIGIR Conference. ACM Press, NY, 232-239.

[24] Porter M. 1981. Snowball: A language for Stemming Algorithms, http://www.snowball.tartarus.org/