MULTILINGUAL QUERYING

Mohammed Mustafa, Hussein Suleman

Department of Computer Science, University of Cape Town , Rondebosch – 7700, Cape Town, Republic of South Africa
mohammed.mustafa@acm.org, hussein@cs.uct.ac.za

Keywords: Multilingual Query, Mixed document, Multilingual Information Retrieval, Language-aware

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, resulting in mixed/multilingual queries with both English and Arabic terms. Current CLIR techniques are optimized for monolingual queries, even if they are translated, but neither mixed-language queries nor searches for mixed-language documents have yet been adequately studied. This paper attempts to address the problem of multilingual querying in CLIR. It shows experimentally that current search engines and IR systems are not language-aware and are not adequate for multilingual querying. The paper then presents the main ingredients that every language-aware solution should take care of.

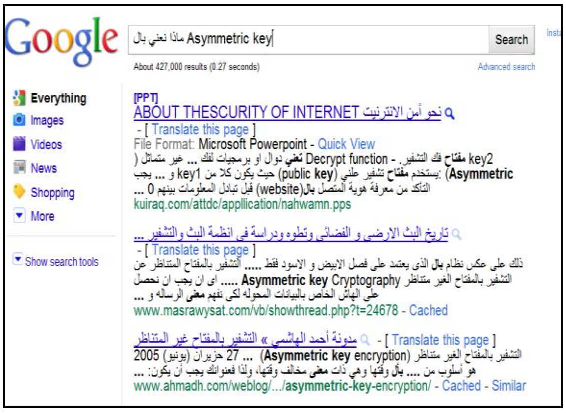
1. INTRODUCTION

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. 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. 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 (Miniwatts Marketing Group, 2011). 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 on 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.

This paper attempts to address the problem of multilingual querying and describes how weighting could be affected by such queries. It focuses on common computer science vocabulary with special attention on Arabic/English bilingual querying. The paper shows experimentally that current search engines are not language-aware systems. It also addresses the main ingredients that every language-aware solution should take care of.



ماذا نعني بال Asymmetric key

About 427,000 results (0.27 seconds)

[PPT]
ABOUT THESCURITY OF INTERNET نحو أمن الانترنيت
- [Translate this page]
File Format: Microsoft Powerpoint - Quick View
key2 مفتاح فك التشفير. - Decrypt function تعني دوال او برمجيات لفك ... غير متماثل (
(Asymmetric :يستخدم مفتاح تشفير علني (public key) حيث يكون كلا من key1 و ... يجب
التأكد من معرفة هوية المتصل بال(website) قبل تبادل المعلومات بينهم 0 ...
kuiraq.com/attdc/appllication/nahwamn.pps

تاريخ البث الارضى و الفضائى وتطوه ودراسة فى انظمة البث والتشفير ...
- [Translate this page]
ذلك على عكس نظام بال الذى يعتمد على فصل الابيض و الاسود فقط ..... التشفير بالمفتاح المتناظر عن
التشفير بالمفتاح الغير متناظر Asymmetric key Cryptography ..... اى ان يجب ان نحصل
على الهاش الخاص بالبيانات المحوله لكى نفهم معنى الرساله و ...
www.masrawysat.com/vb/showthread.php?t=24678 - Cached

مدونة أحمد الهاشمي » التشفير بالمفتاح غير المتناظر - [Translate this page]
التشفير بالمفتاح الغير متناظر (Asymmetric key encryption) ... 27 حزيران (يونيو) 2005
هو أسلوب من .... بال وقتها وهي ذات معنى مخالف وقتها، ولذا فعنوانك يجب أن يكون: ...
www.ahmadh.com/weblog/.../asymmetric-key-encryption/ - Cached - Similar

Figure 1: shows an example of a multilingual query.

2. RELATED WORK

The issue of using bilingual queries and documents has been discussed in the library community. Hansen et al. (2002) enumerated some user requirements for Cross Language Information Retrieval (CLIR) systems, including the support of multilingual queries and the ability to search multiple languages simultaneously. Petrelli et al. (2004) found that the English term in a multilingual query is usually used as a pivot in searching because English is still the dominant language in technical jargon. Rieh and Rieh (2005), in their study of Web searching behaviour, concluded that the querying and searching behaviour is dependent of 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. Lu et al. (2008) tackled the reasons behind using multilingual trends of querying in users' behaviour. The findings, 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 following: using computer technologies, names of magazines and firms; some Chinese words do not have a popular translation; and the culture, such as in Hong Kong, of using both Chinese and English in speaking and writing. Analysis by Lu and his team also showed that there were many queries, which consist of both a Chinese term and its corresponding translation in English. Users in such cases might intend to obtain a higher recall.

CLIR has focused on developing approaches for effective translation of queries (Saralegi, 2010) 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. If the document collection is in more than one language (mixed/multilingual collection), as in text in non-English scientific documents, then the task is that of Multilingual Information Retrieval (MLIR) (Chen and Gey, 2004). Two major architectures for indexing multilingual collections are centralized and distributed (Chen and Gey, 2004; Kishida, 2005). The centralized architecture considers putting all documents, regardless of their languages, into a single centralized index (Nie and Jin, 2003). 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. However, Lin and Chen (2003) showed that unless weights in the centralized approach are adjusted, documents with small collections may be preferred. This is because the number of documents (document frequency) increases while the number of occurrences of a term (term frequency) is unchanged and thus there is overweighting.

A distributed architecture indexes documents in each language separately (Kishida, 2005). Next, the individual ranked lists are merged into a single ranked list. Different merging methods were proposed. Another type of distributed architecture employed putting all documents into a single unified index, as in the centralized approach (Chen and Gey, 2004). Queries are translated into the documents' languages. Next, a monolingual retrieval 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

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

documents is to sum up the scores of these documents (Chen and Gey, 2004). However, there is an explicit assumption in multilingual information retrieval that documents in individual lists do not overlap.

However, none of these CLIR efforts specifically address the problem of language-aware multilingual querying or searching for mixed-language documents. This is because most approaches were designed for monolingual queries or documents.

3. WHY MULTILINGUAL QUERYING

Most languages used in developing countries, including the Arabic world, suffer from a limited modern vocabulary. In addition to the historical backgrounds related to the early days of higher education in non-English countries, the phenomenon of limited vocabulary and multilinguality has three major reasons. First, is the dominance of English in the scientific domain (Miniwatts Marketing Group, 2011). 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. English scientific terms in the Arabic-speaking world, for example, are usually used to simplify ambiguous Arabic scientific terms. This is because most scientific terminology is borrowed from English and it is not always possible to provide precise translations for new terms or to directly express concepts in medicine and technology, for example, in the non-English languages because in most cases the concepts need to be expanded or approximated using context. 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, one of the most significant problems with the Arabicization process is that scientists who perform the process do not usually invite the experts and scientists in a given scientific domain to participate. For instance, the Arabicization of the English terms: ‘brainstorm’, ‘business re-engineering’ and ‘computerization / automation’ are العصف الذهني ، الهندرة and الأتمتة, respectively (The Academy of Arabic Language, 2011). These 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, most of which have their own academy for the development of the language. Each academy Arabicizes new terminology individually, without coordination in most cases with its peers across the Arabic-speaking world (The Academy of Arabic Language, 2011). As a result, scientific modern terms in Gulf countries may be totally different from those in Levantine countries. For instance, the Arabic translation of the scientific English term ‘Deadlock’ has many different dialects on the Web (الجمود، الإقفال، الاستعصاء، التقاطع). In fact, a significant proportion of the Arabic technical terms on the Web were found to be inconsistent and in different regional variants. The problem of regional variants in scientific Arabic terminology grows dramatically with every new term added to the language.

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 in both English and the native languages. 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. Consider the following explanatory example:

Consider a multilingual query *Q* = ‘ مفهوم الـ Inheritance’ (meaning: concept of inheritance) and a document collection consisting of the following six documents:

D$_1$: “مفهوم الـ inheritance يدعم الفكرة الأساسية لإعادة استخدام البرامج”
D$_2$: “مفهوم الوارثة Inheritance يسمح بإنشاء تصنيفات هرمية ”
D$_3$: “The concept of inheritance allows the creation of hierarchical classifications”
D$_4$: “Java does not support the inheritance of multiple superclasses into a subclass. This is different from inheritance in C++. Inheritance in C++..”
D$_5$: “Inheritance is one of the cornerstones of object-oriented programming. Using inheritance you can create a general class that….”
D$_6$: “تؤثر الوراثة بشدة على تعريف المتغيرات. لذلك فإن الوراثة”
Q: مفهوم الـ Inheritance

In this collection, *D*$_2$ and *D*$_3$ are identical, since *D*$_2$ is the exact translation of *D*$_3$. Since *D*$_2$ is in Arabic, the translated English term ‘inheritance’ co-occurs with its Arabic term. This is very common in Arabic scientific writing, especially in references. Table 1 illustrates the document similarity computations when the multilingual query *Q* is submitted to the collection. For simplicity, computations are provided for the keywords: ‘Inheritance’ and ‘مفهوم’only. Similarity is computed in terms of simple TF*IDF. The *DF* and *IDF* for the term ‘inheritance’ is 5 and log(6/5) = 0.07918, respectively, while the *DF* and *IDF* for the term ‘مفهوم’ is 2 and log(6/2) =0.47712, respectively.

According to these computations, the ranking of documents would be D$_1$, D$_2$, D$_4$, D$_5$, D$_3$. It is notable that D$_1$, and D$_2$ have the same scores. Although D$_2$ and D$_3$ are identical, the difference between their scores is disappointing.

The findings also show that dominant documents on the ranked lists are those that contain exactly the same terms in the multilingual query.

Table 1: Computations of ranking.

Docs	inheritance	مفهوم	Documents' scores
	TF * IDF	TF * IDF	
D$_1$	1 * 0.07918	1* 0.47712	0.233913
D$_2$	1* 0.07918	1* 0.47712	0.233913
D$_3$	1* 0.07918	0* 0.47712	0.006270
D$_4$	3* 0.07918	0* 0.47712	0.018808
D$_5$	2* 0.07918	0* 0.47712	0.012539
D$_6$	0* 0.07918	0* 0.47712	0
Q	0.07918	0.47712	-

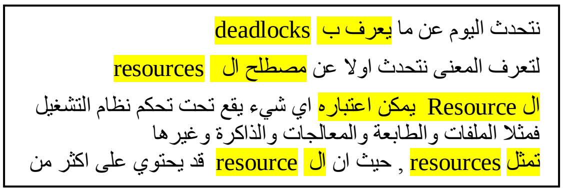
Although D$_4$ is the most highly relevant document, at least in terms of TF, in the collection, it is ranked at the middle of the result list. Also, D$_6$ is not retrieved by the query, although it is highly relevant.

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. ARBIC MULILINGUAL COLLECTIONS

In MLIR the document collection contains at least two languages. Lin and Chen (2003) stated that document collections have two main categories: single language document collections and multilingual document collections. In the single language document collections, all documents are written in a single language. In the second approach, which is the multilingual document collections, documents are written in different languages. Examples include organizations in non-English-speaking countries, which usually have the same content for their websites in different languages. Two types of multilingual data collections are common. The first type consists of several monolingual document collections while the second consists of several monolingual documents plus multilingual documents. A multilingual document is a mixed document that contains different languages.

Scientific non-English documents in Arabic have two distinguishing characteristics that are not found in English documents. Firstly, many multilingual documents contain different terms/portions /snippets/phrases/paragraphs in two languages – usually English is one of them- but in a tightly-integrated manner. This is the worst case of multilingual document in which it contains some terms in English that are strongly and tightly-integrated with Arabic terms in composing sentences, rather than presenting English terms for providing the precise meaning of the Arabic terms, i.e. as a translation. As in multilingual querying, in scientific Arabic documents, you may find a sentence like 'حيث أن الـ deadlock' (meaning: whereas the deadlock) which is a meaningful and strongly integrated sentence, but in two different languages. Obviously, this phrase could be ambiguous or meaningless if we delete the term 'deadlock'. Figure 2 shows a part of a multilingual document taken from the Web in the computer science domain. The document is written in both Arabic and English. It is evident that the text in this document is fully integrated and in two languages. The tightly integrated portions are highlighted.



نتحدث اليوم عن ما يعرف ب deadlocks

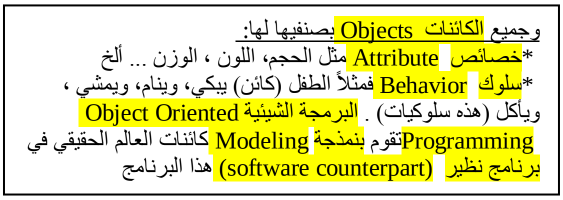
لتعرف المعنى نتحدث اولا عن مصطلح ال resources

ال Resource يمكن اعتباره اي شيء يقع تحت تحكم نظام التشغيل
فمثلا الملفات والطابعة والمعالجات والذاكرة وغيرها
تمثل resources , حيث ان ال resource قد يحتوي على اكثر من

Figure 2 Part of a multilingual document.

Secondly, a considerable number of multilingual documents contain similar description texts/snippets in multiple languages. For instance, in the multilingual phrase '(Hashing) ما هي البعثرة' (meaning: what is the Hashing) the English word (Hashing) is presented as a translation for the Arabic word (البعثرة) and hence removal of the English term will not make the sentence meaningless, unlike in the first case presented in Figure 2. This situation is very common in Arabic Web pages, especially in science, medicine and technology. In fact, formal writing in references and text books usually use such co-occurrences of both Arabic and English scientific terms. Figure 3 shows a part of a scientific Arabic multilingual document taken from the Web in the computer science domain. Most English terms in this document are presented as translations to refine the Arabic terms. This characteristic is also prevalent in other non-English languages. Zhang and Vines (2004) stated that in Chinese Web pages, English terms are very likely to be the translations of their immediately preceding Chinese terms.

Although there is a dominant language in Arabic multilingual documents, the English snippets in them are rich and good candidates for search keys. Moreover, sometimes the same term/word in the same multilingual document is written in different positions but in two different languages, each of which is tightly integrated with its neighbours.



وجميع الكائنات Objects بصنفيها لها:
*خصائص Attribute مثل الحجم، اللون ، الوزن ... ألخ
*سلوك Behavior فمثلاً الطفل (كائن) يبكي، وينام، ويمشي ،
ويأكل (هذه سلوكيات) . البرمجة الشيئية Object Oriented
Programmingتقوم بنمذجة Modeling كائنات العالم الحقيقي في
برنامج نظير (software counterpart) هذا البرنامج

Figure 3 Part of an Arabic multilingual document.

5. EXPERIMENTING WITH CURRENT SEARCH ENGINES

It would be convenient to exploit current search engines to handle multilingual queries. Therefore, a simple experiment was conducted. In this experiment, two international, general and multi-language search engines (Google and Yahoo[2]) were used.

It is known that queries for tests, even if they are simple, should be representative of the queries submitted by users of the target application (Croft, 2009). This approach is followed in this experiment. Hence, to generate queries for the experiment, the selected potential users were a group of 5 students at different academic levels at a computer science department in an Arabic-speaking university. Each potential user in the group was requested to submit examples of 5 queries on common computer science vocabulary. The choice of the query language was deliberately avoided and hence participants could show their natural searching behaviours. Around 25 queries were obtained. All submitted queries were pooled into one set. Duplicates and semi-similar queries were removed. Hence, a cleaned set of 16 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. It is presumed that some students were limited by the modern vocabulary of common computer science in Arabic or, in the best case, they would not want to miss valuable relevant documents due to regional variants. A set of six multilingual queries was selected. The selection of queries was based on a suitable recall: most queries should have suitable relevant documents. Moreover, queries were selected to reflect some of the problematic characteristics of search engines that

[2] http://www.yahoo.com

affect information retrieval, when handling multilingual querying. Thus, all of the selected queries are multilingual. The queries with their translations/meanings in English are listed in Table 2. Queries were numbered (DLIB01-DLIB06) for referencing purposes. The average no. of words per query was found to be 3.3 with 1.3 and 2 as the average number of words for English and Arabic, respectively.

Table 2: Multilingual queries used in the experiment.

Query #	Query	Counterpart in English
DLIB01	Deadlock مفهوم الـ	Concept of deadlock
DLIB02	Secure ماذا نعني بالـ Socket Layer	What is meant by Secure Socket Layer
DLIB03	الفرق بين ال و ال Interpreter Assembler	Difference between Interpreter and Assembler
DLIB04	شرح الـ Polymorphism	Explain Polymorphism
DLIB05	Entity مثال في ال Relationship Model	Entity and Relationship Model, Example
DLIB06	Data تقنيات ال Mining	Data Mining Techniques

It is well-known in IR that the most relevant documents are usually highly desirable by users and should be ranked higher, regardless of the query language(s). Thus, the highly relevant documents are more useful than those that are marginally relevant.

Therefore, in the experiment, the set of six multilingual queries were submitted to both Google and Yahoo and, for each query, the top 15 retrieved documents were examined for their languages and whether they are the most relevant documents. The searches were conducted in January 2011.

Table 3 illustrates results that were obtained from the experiment. For each query listed in Table 2, the majority of the top returned documents is multilingual and contains terms in both Arabic and English that exactly match the query terms, regardless of the ingredient languages of these queries. The reason behind this phenomenon is that current search engines typically lack analysis capabilities in terms of: the mixed and tightly-integrated texts in queries. Due to this limitation of analysis, search engines handle terms in multilingual queries as if they were in a single language.

Table 3: Results using the six queries.

* G=Google, Y=Yahoo

Query No.	No. of mixed docs		No. of English docs		No. of Arabic docs	
	G	Y	G	Y	G	Y
DLIB01	14	15	1	0	0	0
DLIB02	15	15	0	0	0	0
DLIB03	13	15	2	0	0	0
DLIB04	15	15	0	0	0	0
DLIB05	14	15	1	0	0	0
DLIB06	13	15	2	0	0	0

In the results, monolingual Arabic documents did not appear in the top 15 documents for all queries. This is because monolingual scientific Arabic documents are very rare, at least in terms of common computer science. Yahoo did not retrieve any monolingual document in English at the top 15 documents while Google did. When the top 15 documents were explored by 3 staff members in computer science, it was noted that most of them are relatively poor in terms of relevance. They did not provide much information. In contrast, there were highly relevant documents, mostly in English, that were ranked lower - and many times much lower - in the result list. Hence, a lot of excellent documents, containing rich information, could be easily missed by users. It was noticed that some Persian documents had been mistakenly retrieved because of the shared script between Arabic and Persian. However, it is clear from the experiment that mixed pages in different languages on common computer science are relatively few compared with documents in English, at least in terms of discovery by a search engine.

Since multilingual documents might be retrieved by a monolingual Arabic query because the majority of scientific Arabic documents are mixed, one might ask: why not issue a monolingual Arabic query. In addition to the avoidance of regional variation and the non-availability of possible scientific terms in Arabic, the answer is simply because many scientific Arabic terms are shared with words in the common literacy in Arabic.

From this simple experiment, it is concluded that current search engines cannot yet handle multilingual queries and cannot guarantee that the top ranked documents are the best ones, depending on their term frequencies and document frequencies.

It is possible to say that current search engines are language-unaware IR systems.

6. LANGUAGE-AWARE SOLUTIONS

Ideally, language-aware solutions would have the ability to match multilingual terms of queries with monolingual/multilingual documents and vice versa. One of the major limitations in current approaches - when multilingual querying is considered - is that they handle terms in these multilingual queries as if they were presented in a single language, and consequently the same weighting scheme would be applied to all terms regardless of their languages, hence resulting in typical matching of terms. When it comes to multilingual querying, it may be necessary to assign, using statistical methods, some reasonable weights to terms in different languages in multilingual queries, so as not to favour one language with respect to another. In fact, the significance of different portions in multilingual queries is different. Usually English terms in multilingual queries are key search terms and useful clues. This is shown in multilingual queries in Table 2. Such modified weighting would, at least, make both monolingual and multilingual documents comparable. It is also necessary to account for mixed phrases that are tightly-integrated and those phrases that co-occur for simplicity purposes. It is also essential that language-aware weighting should realize that in non-English scientific documents the same term may appear in two different languages in the same multilingual document. In this explanatory example above, it may be convenient to re-weight terms in a multilingual query across languages and thus make document scores comparable, regardless of their languages. Dependent language identification algorithms may be needed for such a process of language identification, depending on the collection. It may be necessary to mark each term/word/portion/paragraph with its language during the indexing process of documents. Moreover, appropriate standard IR algorithms such as stemming, stopwords and morphological analyses are needed for language-aware solutions. Implementation of such guidelines may make it possible to enhance querying in multiple languages by interpreting and handling these queries as language-specific instead of language-independent.

Both centralized and distributed architectures were designed for indexing several monolingual documents, rather than documents with two languages, as is explained in next section. Therefore, language-aware solutions may need to adopt other techniques for indexing and storing mixed (both monolingual and multilingual) document collections.

6.1 Centralized Index and Multilingual Querying

It is known that a centralized architecture overweights documents in small collection because the total number of documents increases (Lin and Chen, 2003), resulting in degrade IR performance, especially in scientific collections. In such collections, it is expected that the English collection would be much bigger than collections in non-English languages. However, Lin and Chen (2003) did not consider the number of occurrences of a term in different languages, as in multilingual documents. In such queries and documents both the number of occurrences of a term (TF) and the number of documents increase. Moreover, the centralized approach in multilingual documents does not take into account the difference between English terms that are tightly-integrated inside the Arabic texts, as in the phrase ‘ مفهوم الـ Class’ (meaning: the concept of class), and the terms that are placed in a document as a translation to explain/approximate non-English scientific terms, as in the multilingual phrase ‘Class فئة’. The difference between these two scenarios could affect the final score. This is clear in scores of D_2 and D_3 in the explanatory example above. This is because the scientific terms in the Arabic document would be computed twice and independently from its translations, although both the documents are identical and some of their terms are similar but in different languages, as in ‘inheritance’ and ‘الوراثة’. Thus, the centralized index cannot guarantee that the top ranked documents are the best ones (most relevant documents).

6.2 Distributed Index and Multilingual Querying

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. 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 a multilingual query will appear in these partitioned multilingual documents.

The second option that could be applied for multilingual documents in a distributed approach is to implement the second version of the distributed architecture, which puts all documents in a single index, as in a centralized index. At first look this method sounds more adequate for multilingual documents. But such documents may overlap in individual lists. Thus, multilingual documents may be ranked at the top if we sum up their scores across individual lists.

7. CONCLUSIONS AND FUTURE WORK

Non-English-speaking users, such as Arabic speakers, are not able to express terminology in their native languages. Therefore, their queries are usually expressed in a multilingual form. It may no longer be possible to constrain users, especially non-English-speaking ones, to a single language when searching for scientific documents. CLIR has focused on developing approaches for effective translation of queries but neither mixed-language queries nor searches for mixed-language documents have yet been adequately studied. Most current search engines and traditional CLIR systems perform poorly when handling multilingual querying. Current methods to index multilingual collections might not be the optimal solutions. Therefore, the paper argues that there is a need for multilingual querying. It showed experimentally that current search engines and CLIR systems are inadequate to handle multilingual queries. The paper addresses also the potential components for building language-aware solutions.

Most corpora are built from news, legal documents and encyclopedias (Croft et. al, 2009) and they usually contain several monolingual documents in different languages, with each collection in a given language. Therefore, currently, the authors are developing a multilingual corpus of common computer science vocabulary. One of the major components also in the future work is developing weighting algorithms. It may be necessary to assign reasonable weights to terms in different languages in multilingual queries, so as not to favour one language with respect to another. New merging methods, which have the ability to handle multilingualism, may be investigated.

8. REFERENCES

Chen, A., Gey F., 2004. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding, *Journal of Information Retrieval*, Kluwer Academic Publishers, *Volume* (7), pp.149-182

Croft , B., Metzler, D., Strohman, T, 2009. *Search Engines: Information Retrieval in Practice*, Addison-Wesley, USA, 1

Hansen, P., Petrelli, D., Karlgren, J., Beaulieu, M., Sanderson, M., 2002. User-centered interface design for cross-language information retrieval, *Proceedings of the twenty-fifth annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Pres, New York, pp. 383–384

Kishida, K., 2005. Technical issues of cross-language information retrieval: a review, *Journal of Information Processing and Management*, Elsevier B. V., Volume (41), pp.433 - 455

Lin, W.,Chen, H., 2003. Merging Mechanisms in Multilingual Information Retrieval, *Advances in Cross-Language Information Retrieval LNCS*, Springer-Verlag, Volume (2785), pp . 175-186

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*, Milwaukee, Wisconsin, USA

Miniwatts Marketing Group (2011), "Internet World Stats Usage and Population Statistics", Available at: http://www.internetworldstats.com/, Last accessed 20 -2- 2011

Nie, J. Y., Jin, F., 2003. A multilingual approach to multilingual retrieval. In C. Peters, et al., eds. *Advances in cross-language information retrieval LNCS* Springer-Verlag, Berlin, 2785, pp. 101–110

Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P., Hansen, P., 2004. Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system, *Journal of the American Society for Information Science and Technology*, Volume (55), pp. 923–934

Rieh, H., Rieh, S., 2005. Web Search across Languages: Preference and Behavior of Bilingual Academic Users in Korea, *Library & Information Science Research*, Volume (27), pp. 249-263

Saralegi, X. and Lopez de Lacalle, M., 2010. Estimating Translation Probabilities from the Web for Structured Queries on CLIR, *In Proceedings of the 32nd European Conference on Information Retrieval (ECIR)*, Springer-Verlag, Berlin, LNCS (5993), pp. 586–589

The Academy of Arabic Language, 2011. Sudan Office

Zhang, Y., Vines, P., 2004. Detection and translation of oov terms prior to query time. In *SIGIR '04*, ACM Press,UK, pp.524-525