Evaluating Google queries based on language preferences

Ahmed F. Al-Eroud, Mohammad A. Al-Ramahi, Mohammed N. Al-Kabi, Izzat M. Alsmadi and Emad M. Al-Shawakfa
Faculty of Information Technology, Yarmouk University, Jordan

Abstract

This paper evaluates the assumption that users expect search engines to retrieve the same results for queries regardless of the language or the location of the originator. The dependency of the Google search engine on the language and location from which the query is submitted has been evaluated. The most popular queries in Arabic language were selected and translated into English for comparison using the Google translator. When studying keyword traffic on both Google search based keyword tool and Google Insights for Search, results showed that 67% of the Arab Internet users prefer to use English queries instead of their Arabic counterpart. When studying Google responses to some popular queries we have found that Google ranking algorithm depends on the language of the query more than on the keyword popularity. Although results justify search engines' favouritism of giving documents in English priority over those of other languages, nonetheless, future search engine indexers should separate the document language from its content in a structure that makes the language a pluggable attribute for those indexed documents.

Keywords

Cross-Lingual Information Retrieval (CLIR); information retrieval; page ranking; query processing and indexing; search engines

1. Introduction

Google have recently launched 12 new local domains for some Arabic countries. Users of languages other than English can either search using English keywords or search using their own native language. Retrieved search results of non-English queries in most cases may not exactly match the search results of the English queries. This can be justified through saying that the users who search using a keyword in a particular language are interested first in getting results in that specific language. Users who also search from a specific location may want to get first results from their own country or area they are in as being more relevant pages than those of other languages or from other continents. However, in both cases, results should eventually be comparable. This means that indexes or the libraries of search engines should isolate the layer of the location and the language from the actual content and documents retrieved and indexed in their own library or database.

In this research, we have evaluated some of the issues related to information retrieved from different perspectives (languages and structure). The ultimate goal is to propose building indexers that are language and structure independent. By structure independent, we mean that search engines should be smart enough to know the close words (i.e. using thesaurus in some cases) in the native language of the user. First matches should always be of those that exactly match the query terms with the listed result. However, eventually similar or comparable words should always be part of the retrieved results. This is necessary, since the translations of certain phrases by different human translators may not exactly match. Many popularity tools are used by Google to measure keyword popularities, and the traffic on each keyword. Google Insights for Search and Search-Based Keyword tool (i.e. SBK tool) are some examples of the important tools used by Google to measure the search traffic on keywords. In this research the SBK tool has been used to measure the search traffic of different queries and the translated ones in order to examine whether Google takes the traffic issue into their ranking algorithm or not [1].

Corresponding author:
Izzat Alsmadi, Department of Computer Information Systems, Faculty of Information Technology and Computer Science, Yarmouk University, Irbid 21163, Jordan.
Email: ialsmadi@yu.edu.jo

2. Related work

There are several papers and researchers that evaluated the effectiveness of search engines and the relevance of the retrieved results. In addition there are several researchers who have studied multilingual information retrieval systems.

Jeongwoo et al. studied how to present multiple language answers probabilistically for questions that are independent of the language. Logistic regression was used to estimate the probability of candidate-suggested Chinese and Japanese answers to questions to re-rank answers within the answers set. This method has improved the answering performance by 40% for Chinese questions, and by 45% for Japanese questions [2].

Guo et al. proposed a multi-lingual information retrieval system for patent documents in English and Japanese languages; different web translators were used to translate queries such as Google and Excite translators. The language-independent indexing technology was used to process the text collections in various Asian languages. The results have indicated that the proposed method has achieved effective results; however, the proposed system was not a web-based one. In addition, no relevance feedback procedure was used [3].

The contextual information during the query session was exploited by Huang et al. to suggest appropriate queries. The proposed terms are produced according to their relevance with query session. Huang et al. study tried to infer the user information needs by analysing session term relevance of each user from the collected logs of user queries in web servers. Clustering was used to collect suitable suggested terms for queries. Results have showed that the proposed technique significantly helps the users to select the appropriate terms [4].

Fonseca et al. studied the effects of adopting association rules used mainly in data mining to get users' suggested queries. A web log of user queries that exceeds 2.3 million queries was used to test the adopted method. The results of this study showed that the percentage of queries correctly suggested within the five top related terms was 90.5%. This percentage is 93.4% for randomly selected queries from the log of user queries. On the other hand, this percentage is 92.2% for choosing the right query given to the user from the list of suggested queries. Fonseca et al. showed that adopting this method for query expansion will help to improve search engine answers by 23.1% [5].

Xinhui et al. adopted topic-relevancy to expand queries within four steps process. In the first step the Chinese short terms were extracted automatically from the documents set in order to build an index. The second step was dedicated to searches that were based on these extracted short terms from both the queries and documents' set. The third step was to acquire automatically topic-relevant terms from Google search engine for each short query term. This step was based on topic-relevant terms and relevant terms from the top 30 initially retrieved documents to expand queries. The fourth step was used to perform a search for the second time using the expanded queries to get new ranking. The results of that study showed that this method is better than standard Rocchio expansion [6].

Al-Maskari et al. conducted an experimental study to evaluate the efficiency and effectiveness of Google in processing Arabic queries. The study was based on using 26 Arabic users, where each user was asked to choose four different Arabic queries, with a total of 104 Arabic queries. The search was limited to four subjects: religion, art, health, and politics. Users were asked to save five relevant pages within 12 minutes for each of the above four subjects mentioned before, besides scoring the relevancy of each of the 10 pages resulting from issuing an Arabic query to Google. The scores were limited to three, highly relevant, reasonably relevant, and not relevant. The users also performed measurements concerning satisfaction, accuracy, coverage, and ranking. The overall results show a dissatisfaction of Google users who were searching for Arabic web pages [7].

A system to answer crime-related queries in terms of spatial information was proposed by Chengyang et al. The system was applied and tested using information related to crimes in Texas. Such information can be used by law enforcement agents for laboratory or real-time investigations. The model combined natural language processing techniques such as speech tagging and semantic parsing with schema matching. The experimental results showed that an approach that relies on natural language processing techniques applied on the user input can be practically and effectively used [8]. Janevski et al.'s study proposed a web search portal to deal with the Macedonian language in particular. All non-English languages may suffer from the fact that English is the dominant language in information retrieval and the Internet. Each language may have its own structure and way of forming words, nouns, verbs and statements. Authors found out that stemming words in Macedonian may have different roles than those in English or other languages. A major goal of such language oriented web search portals is to enable language users to use them and retrieve relevant information, similar to the widely internationally known English websites [9].

Chew et al. studied the effects of language relatedness on the performance of cross language information retrieval systems [10]. The importance of that approach was to measure the effects of including a Semitic language on cross-language information retrieval (i.e. Arabic). The results indicated that adding linguistic pre-processing to CLIR (cross language information retrieval) has enhanced the performance; however, adding more languages was generally a beneficial step.

Airio discussed the effect of using dictionary and translation on web queries. Results showed that original language retrieved results may not be identical to the translated one. However, the study justifies logically the usage of dictionary or web translation on web queries [11].

Dolamic et al.'s paper evaluates an information retrieval model applied to a collection of documents written in different languages. The paper uses several different languages besides English: Chinese, French, German etc., especially when the query language is different from the document language [12].

Singh et al. evaluated the language effect on Google searches for a particular health information. The paper found out that in this field the majority of papers are actually written in English, although they were originated from countries of non-English speakers. Authors have also found that one reason for the linguistic digital divide is that the majority of health and food web pages are not translated into multiple languages and/or that their cross-language retrieval by search engines is poor [13].

Methods involving the usage of machine translation for CLIR, parallel corpora and machine readable bilingual dictionaries have all been tested in many studies with a varying degree of success. Ballesteros et al. studied combined corpus analysis techniques with query expansion to disambiguate terms and phrases. The paper showed that this combination can reduce the error associated with query translation [14].

Lee et al. presented a method to resolve ambiguity in CLIR using dynamic incremental clustering. Their paper presented the machine translation from English into Korean and Japanese. They evaluated this method on TREC-6 and TREC-8, and proved that it can greatly resolve ambiguity compared with other methods [15].

Gonzalo et al. presented an approach to CLIR based on the EuroWorldNet (EWN) multilingual semantic database. EWN is based on WordNet database with basic semantic relations between words for diverse European languages. Both documents and queries written in any language within the database are indexed in a space of language-independent concepts, and searches followed term weighting and matching processes [16].

Another paper that presented a CLR technique is that of Hermes et al. This paper used the method for retrieving medical information. The method combines query terms and related medical concepts, where the results of the CLIR method are compared with those of standard space model algorithms and showed comparable results [17].

Zhuhadar et al. proposed Multi-Lingual Information Retrieval (MLIR) approach of domain-specific information retrieval (DSIR), which is e-learning. The approach followed was a synergistic one, which was between the thesaurus-based approach and the corpus-based approach. A simple bilingual listing of terms, phrases, concepts, and subconcepts, as well as hierarchical structure of the ontology was used to define the relationship between concepts/subconcepts. A term vector translation approach was involved for translation, where the goal was to map statistical information about term usage between languages by using techniques map sets of *tf-idf* term weights from English to Spanish [18].

The support for multiple languages is not available in most current search engines [19]. A toolkit that allows users to build, access, and maintain multiple document collections in multiple languages was built by M. Chau et al. According to the authors such a tool will be very useful for vertical search engine development.

Efthimis et al. have investigated how search engines respond to non-English queries and more specifically to Greek language queries. A set of navigational queries for known Greek organizations was created, and searched on a set of Global and Greek search engines. The analysis showed that the global search engines ignore the characteristics of the Greek language, hence treating Greek queries differently. Although Google was the best global search engine in handling Greek queries, it was able to find the correct answer to only 73.91% of the English and 60.37% of the Greek queries. The search engines seem to have poor coverage of the Greek web pages [20].

Moukdad compared the performance of three general search engines that were specifically designed to handle the linguistic characteristics of Arabic. The limitations of general search engines (i.e. Google, AltaVista) in retrieving Arabic documents was clear, a high number of documents were lost by such search engines where only the exact forms of Arabic words are considered in the retrieval process [21].

3. Goals and approaches

This study evaluated the differences in results retrieved from Google search engine for a total of 1100 Arabic popular queries and their English translations. Figure 1 shows the methodology of this study. Our queries set were collected using the Google suggested list; this list suggests at most 10 popular queries for a particular letter(s). There are 28 letters in Arabic; however, some letters have more than one shape, such as the first letter in the Arabic alphabets, Alif, (ى, أ,ا,آ,إ), besides Waaw (ؤ, و) and Yaa' (ئ, ي), so the total number of Arabic letters used in query collection step was 34. The Google suggested automatically a list of a maximum of 10 auto suggested queries. In the query collection step we combined each letter of the 34 letters with the other 33 to build the Google query database, so for each letter the total number of collected queries was (33 × 10) which is 330 queries, so we collected (330 × 34), which was more than 11,000 queries. The Google.com domain was used in our research, which is considered as a global domain and not tied to any particular country. Our queries were collected in 11 days, during the same time of year in 2009 and 2010. The goal of collecting such large number of queries in this period was to be more consistent in deciding Arabic users' preferences; the similarity between our collected queries in these two periods was about 72% for all Arabic letters (see the Appendix).

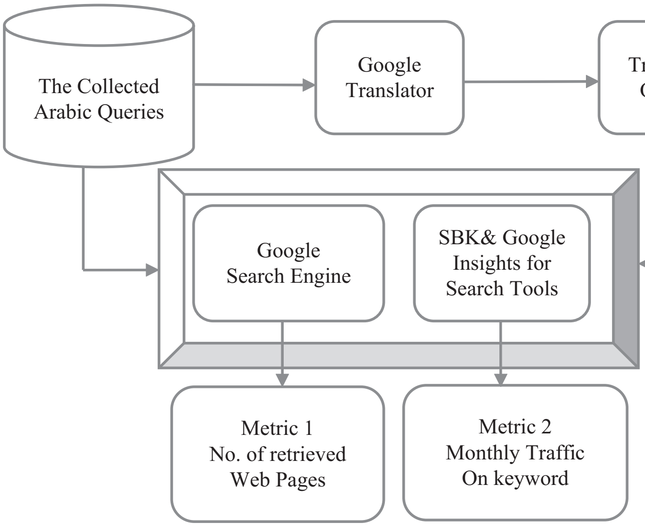

The Collected Arabic Queries
Google Translator
Translated Queries
Google Search Engine
SBK& Google Insights for Search Tools
Metric 1 No. of retrieved Web Pages
Metric 2 Monthly Traffic On keyword

Figure 1. The workflow of the study.

The first step in our approach is the translation of the collected Arabic queries. For consistency Google translator was used in the translation process. In the second step, the lists of the search results for the Arabic queries and their English translations are collected, in order to discover whether these two lists of the search results are correlated. In the third step, Search-Based Keyword tool (SBK tool) is used to see the monthly traffic on the collected queries. The Arabic queries are selected according to traffic they received and hence considered popular. Queries in Arabic and English are entered to the SBK tool to see their monthly traffic. Our goal from using this tool is to examine whether Arabic Internet surfers prefer using Arabic or English language in their queries. Google can benefit from this in deciding whether to process the query in its original language or to process it in a language in which the keyword receive higher traffic. To examine the popularity of each keyword Google Insights for Search popularity estimator is also used to see whether the term is used more often in Arabic or in English, Google Insights for Search statistics gives an indication of Arabic users' behaviour in formulating their queries.

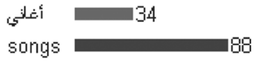
4. Experiment and evaluation

In our experiments two metrics have been used: the number of retrieved pages and the monthly traffic on particular keywords. The number of retrieved pages metric was evaluated by submitting all queries and getting the total number of pages retrieved by Google search engine for each word and its translation.

The estimated monthly traffic on keywords was collected by submitting each query to SBK; the final experiment was made using Google Insights for Search statistics to measure the popularity of particular keywords on Arabic countries.

Google Insights for Search statistics was used to measure the consistency with SBK tool. Google Insights for Search gives an indication of term popularity in different regions of the world; we found that Google Insights for Search popularities was consistent with those of SBK tool. Each query of our selected ones was submitted in Google Insights for Search page; the translated English query was also submitted to see which query was more popular. Table 1 shows how we used Google Insights for Search statistics in our experiment. Each selected Arabic query as well as its translation was submitted

Table 1. Query: ‘اغاني’, songs, popularity



Filters	Query	Translation
	أغاني	Songs
Search type	web search	
Year	2009	
Region	Egypt	
Search category	All categories	

Source: Google Insights for Search.

Table 2. Normalized search traffic on the query 'اغاني', songs in some Arab countries

Region	اغاني	Songs	Ratio between Arabic and English queries
Syria	100	6	0.06
Libya	92	4	0.04
Oman	66	21	0.31
Palestinian Territory	52	3	0.05
Yemen	51	6	0.11
Sudan	41	12	0.29
Jordan	38	6	0.15
Egypt	14	5	0.35

Source: Google Insights for Search.

to see their popularity (lines with number after each one). Table 1 represents the percentage of using the term 'songs', اغاني, and its Arabic translation in Google search engine, given that the number beside the terms in Table 1 does not indicate the actual total hits on each query as such numbers are normalized by Google.

The setting used on Google Insights for Search experiment is summarized in Table 1; first of all, the year 2009 was used as our reference year for measuring popularity of each term. In addition we used the largest Arab country which is Egypt in the Google Insights for Search statistics. The normalized search popularity for terms are drilled down in Table 2.

Table 3 shows the traffic statistics for the single term query 'Yahoo' (ياهو). The table shows the normalized traffic on the English term and its translation in Arabic. It appears according to Google Insights that most people in Arab countries prefer to use the English single term query for the keyword 'Yahoo' instead of its Arabic translation 'Yahoo' (ياهو), the overall ratio of using Arabic to English query for the query 'Yahoo' is about 13%.

4.1. Number of retrieved pages

In this section, we evaluate the results set retrieved by Google for an Arabic query and the result set in English for the same query after translating it using Google translator. Table 4 shows a sample of the collected queries. The queries used in

Table 3. Normalized search traffic on the query 'ياهو' (Yahoo) in some Arab countries

Region	ياهو	Yahoo	Ratio between Arabic and English query
Syria	11	75	0.14
Libya	9	79	0.11
Oman	7	80	0.08
Palestinian Territory	14	65	0.21
Yemen	9	72	0.12
Sudan	4	74	0.05
Jordan	8	76	0.10
Egypt	17	86	0.19

Source: Google Insights for Search.

Table 4. Number of retrieved pages for 11 Arabic queries and their English translations

i	Google results	English query	Google results	Arabic query
1	1,030,000,000	Games	99,700,000	العاب
2	85,500,000	Girls Games	17,800,000	العاب بنات
3	174,000,000	Car Games	5,210,000	العاب سيارات
4	212,000,000	Sun Games	2,360,000	العاب شمس
5	35,500,000	Cooking Games	5,380,000	العاب طبخ
6	176,000,000	Flash Games	12,700,000	العاب فلاش
7	304,000,000	Songs	47,800,000	اغاني
8	897,000,000	Youtube	4,330,000	اليوتيوب
9	1,620,000	Nasheed	12,000,000	اناشيد
10	13,000,000	Melodies	2,890,000	الحان
11	11,400,000	Appolo 11	20,100,000	أبولو 11

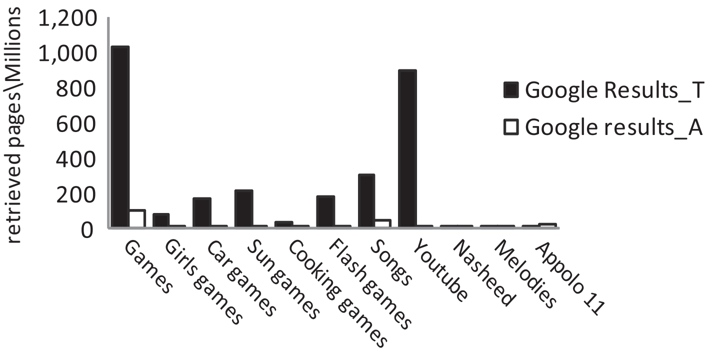


Figure 2. Retrieved pages for Arabic/English queries.

evaluation are distributed on 28 Arabic letters, in addition to different shapes of 'Alif' (ى, أ,ا,آ,إ), besides 'Waaw' (ؤ, و) and 'Yaa' (ئ, ي). When numbers of retrieved pages for the queries in Table 4 are converted into bar chart in Figure 2, we see that the number of retrieved pages for the translated queries is the dominant for the majority of queries. For example when the single term English query YouTube is entered, 897,000,000 pages were retrieved for the translated query (YouTube) in English, compared with only 4,330,000 pages when the term 'YouTube' (اليوتيوب) in Arabic was submitted.

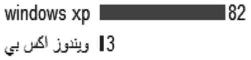
Google ranked the related Microsoft page in the fifth place when the Arabic query, 'ويندوز اكس بي' (Windows XP) was submitted, while the first page retrieved when the English query 'Windows XP' was submitted, which is more popular in terms of the number of retrieved pages and according to traffic statistics on Google Insights statistics is the Microsoft homepage (windows.microsoft.com/enau/windows/.../windows-XP/) page.

In this case we believe that the related Microsoft web page (http://windows.microsoft.com/./windows/windows-XP/) will be more suitable to be number one on the list of search results rendered by Google. When the same query (Windows XP) was analysed in Google Insights for Search website inside the largest Arabic country which is Egypt, it was found that extremely all users in Egypt used the term 'Windows XP' in English instead of 'ويندوز اكس بي' (Windows XP) in Arabic. This indicates that many users who submit 'Windows XP' query are interested in those pages that are also in English. The Google Insights for Search statistics for the query windows XP in both Arabic and English are summarized in Table 5. Only 18,600,000 pages were retrieved with the query 'ويندوز اكس بي' (Windows XP) compared with 211,000,000 when the query 'Windows XP' was submitted; these numbers may be justifiable, due to the large number of English indexed web pages; however for the query 'ويندوز اكس بي' (Windows XP), such a list of retrieved pages may miss many important English pages, although about 96% (82/85) of Arabic users use the phrase Windows XP (in English).

4.2. The estimated traffic on keywords

The second metric used in our experiment is the estimated traffic on each keyword of the selected queries. The metric is measured on the set of the collected keywords from the selected queries to estimate the traffic on each Arabic query (the

Table 5. Query (Windows XP, 'ويندوز اكس بي') popularity


windows xp 82
ويندوز اكس بي 3

	Query	Translation
	ويندوز اكس بي	Windows XP
Number of retrieved pages	18,600,000	211,000,000
Search type	web search	
Year	2009	
Region	Egypt	
Search category	All categories	

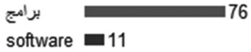
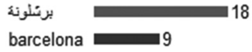
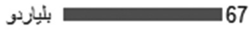
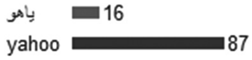
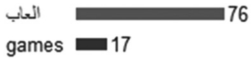
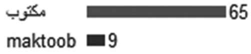
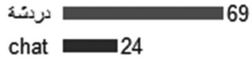
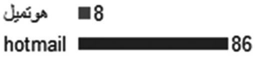
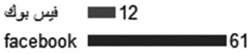
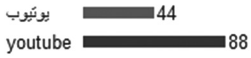
Source: Google Insights for Search.

accumulated monthly hits on such keywords). The goal of finding the search traffic on each query is to decide whether there is a relationship between the search traffic on a particular query and the language in which the query is frequently submitted. The evaluated queries consisted of our collected queries and their corresponding translations in English. The set of selected queries was chosen from the Google auto-suggestion list (which is proportional to the keyword traffic). The queries were selected by submitting each of the 28 letters (as well the different shapes of first Arabic alphabet Alif, and two different shapes (Waaw, و, ؤ and Yaa’, ئ, ي) and then capturing the most popular traffic from the list in which Google auto-suggests at most 10 popular queries. In the queries collection step we also combined each letter of the 34 letters with the other 33 to build the Google query database, so for each letter the total number of collected queries was (33*10) which is 330, so we collected (330*34), about 11000 queries.

To solve the local Google domains problem, the Google.com domain is used instead of selecting a particular domain from the set of 12 Arabic domains of Google, where the monthly traffic on each keyword is retrieved by using the Search-Based Keyword tool to find the monthly traffic received by our selected queries. Table 6 shows a sample of the selected queries and their translation. The first column represents the query used, the second column is the monthly traffic on the query using Search-Based Keyword tool, the third column represents the query translation in which Google translator is used for more consistency, and finally the fourth column represents the monthly traffic on the query after translation. The measured traffic by Search-Based Keyword tool is selected to be on Arabic countries only, hence, all Arabic countries that SBK tool support are selected in the experiment setting. The measured monthly traffic represents the total number of hits on each keyword in the selected countries. For some particular queries the monthly traffic on Arabic keywords was greater.

For example, the monthly traffic on the Arabic keyword ‘Maktoob’ (مكتوب) is 863,650 hits compared with only 67,316 hits on the translated English term Maktoob, which represents only 0.07% of the total monthly traffic on this keyword. For some other queries the monthly traffic on the English translation of a particular keyword is greater than the traffic on the same keyword in Arabic; for the Arabic term ‘Yahoo’ (ياهو) only 201,617 hits are submitted compared with 5,700,967 hits when the English term Yahoo is submitted to Google search engine. This represents only 0.03% of the traffic on this keyword. For some queries we can find some consistency between the traffic on the Arabic keyword and that for its English Translation. For example, the monthly traffic on the Arabic term ‘Chat’ (دردشة) is 65,529. On the same row the translated

Table 6. The number of retrieved pages for a set of Arabic queries and their translation

Google Insights popularity (normalized)	SBK* results	Translated English query	SBK* results	Arabic query
	213,005	Software	702,503	برامج
	170,007	Barcelona	102,601	برشلونة
	573,922	Billiard	114,118	بلياردو
	5,700,967	Yahoo	201,617	ياهو
	2,968,500	Games	100,718	العاب
	67,316	Maktoob	863,650	مكتوب
	43,328	Chat	65,529	دردشة
	2,395,432	Hotmail	144,810	هوتميل
	1,597,451	Facebook	479,421	فيس بوك
	3,845,075	YouTube	61,256	يوتيوب

*SBK: Search-Based Keyword tool.

English term ‘chat’ is submitted monthly about 43,328 times. This represents about 66% of the total monthly traffic on this keyword.

4.3. Google Insights for Search statistics

The statistics on the last column in the traffic table (Table 6) is extracted from Google Insights for Search page, which is used by Google to predict the traffic trend of a particular query; in addition it is used to compare the traffic on a particular keyword during a specific period of time. On the Google Insights for Search popularity column on Table 6, lines with different highlights are drawn, where the length of each line represents the particular keyword popularity in the Google database and how often it is used by Arabic users in their search. We chose a period of one year to measure specific keyword popularity. Egypt, which is the largest country in the Arab world, was selected to collect Google Insights for Search statistics. From Table 7 we notice that Arabic user preferences are divided into three clusters. The first cluster consists of English popular keywords, where keywords with high traffic are called popular. In such terms our findings show that only 23.8% of Arabic users in Egypt prefer using such terms in Arabic. An example of such keywords is the term Facebook, which is a popular term in English. The percentage of the Arabic users who uses the synonym (فيس بوك) is only about 19% of the total Arabic users (searching for this term). The second cluster consists of Arabic popular terms; the popularity of the term in Arabic is the name itself not only the traffic on such keywords. For such Arabic terms our findings show that Arabic users prefer to use the original Arabic popular term instead of its English translations; about 86.4% of users in Egypt use such Arabic terms instead of their English translations. Arabic terms such as *Maktoob* (مكتوب) and *Ajeeb* (عجيب) are popular terms in this cluster. The terms in third cluster are those Arabic keywords that have high traffic. The terms in this category are called ‘popular’ because of the high traffic on them, although the traffic on the Arabic terms on this cluster is about 59% of total traffic on the total number of keywords. The English terms in this cluster receive a comparable traffic of about 41%.

5. Google and language preferences

The goal of this paper was to determine whether Google takes into account the traffic on specific keywords as well as their translations, their synonyms (even if they are in different languages) and their related words. To be specific the focus of this research is to assess Google’s ability to take the traffic on specific keywords and their translations into English into account; if so, then this shall be reflected in the ranking of Google search result pages. From our statistics above, we have chosen two sets of queries that represent the three clusters mentioned earlier. Table 8 shows

Table 7. Traffic table for Arabic queries distributed on three clusters

Keyword cluster	Average traffic on translated Arabic query	Average traffic on translated English query	Average traffic on original query	Sample queries
Popular English queries*	23.8%	—	76.2%	Facebook, Yahoo Hotmail, Twitter
Popular Arabic queries*	—	13.6%	86.4%	Maktoob, Alrai Arab news, Ajeeb
Arabic queries with high traffic**	—	40.7%	59.3%	Games, Songs, Software

* Popular queries due to the name itself.
** Not popular names but have high traffic.

Table 8. Google rank for English query ‘Windows 7’ and Arabic translation ‘7 ويندوز’

Rank	Transliterated Arabic query ‘7 ويندوز’	Original English query ‘Windows 7’
1	www.m7shsh.com/vb/116288.html	www.microsoft.com/windows/windows-7/default.aspx
2	www.libyanyouths.com/vb/t12070.html	ar.wikipedia.org/wiki/7_ويندوز
3	ar.wikipedia.org/wiki/7_ ويندوز	en.wikipedia.org/wiki/Windows_7
4	www.vip600.com/7/	en.wikipedia.org/wiki/Windows_7
5	windows.microsoft.com/arxm/windows7/.../home?os...	www.libyanyouths.com/vb/t12070.html

Table 9. Popular English queries

High traffic keyword (first set)	Low traffic keywords (second set)
Kaspersky	كاسبر سكاي
Windows 7	7 ويندوز
Acer	ايسر
TOEFL	توفل
Java	جافا
…..	……

Google rank for the phrase ‘Windows 7’ and its Arabic translation. Table 9 shows some of the popular English words with their Arabic translation (which is not a real translation; it is a transliteration of the original English word) and their Arabic transliteration

If we examine the first set of answers when the English query ‘Windows 7’, which has high traffic, is submitted, and then compare this set with the set of answers of the transliterated Arabic query ‘7 ويندوز’, with less traffic, we find that, although the first result when the transliterated Arabic query (Windows 7, 7 ويندوز) is the one which has the highest rank by Google, unfortunately, the first retrieved page is not the best one. Microsoft website, which is the most important page, which should be retrieved and ranked first, has the rank number five.

On the other hand when the English query ‘Windows 7’ is submitted, the first result is the Microsoft one, which is actually the most appropriate result for such query. We conclude from the above statistics that, if any keyword is popular in a particular language, then the first result for any user query that contains that keyword or its translation should be in that language, even, if the query is in a different language. From our experiments above, Google depends on the language in which the query is submitted; for the majority of queries the first answer is always in the query language even if the query in that language is not popular.

6. Conclusion

In this paper we have evaluated two metrics on Google search engine. The studied metrics are the number of retrieved pages for a particular query and the monthly traffic on the same query. The relationship between these metrics and the language of the top results pages retrieved by Google is analysed to examine whether Google takes into account the number of retrieved pages and the traffic on a particular query when retrieving relevant pages for that query, even if such relevant pages are not in the language in which the query is submitted. A set of popular queries distributed on 34 Arabic characters were selected, translated and then submitted to Google search page. Google Translator was used to translate Arabic queries into English.

The first experiment was to examine whether the Arabic popular queries would retrieve more pages in Arabic. Afterwards the queries were translated into English and submitted on the Google search page. The results indicated that 60% of the queries retrieved more pages in English, although most of them were popular Arabic queries. Nonetheless, such results could be considered normal due to the large number of indexed English web pages. If an Arabic query is submitted in Arabic and there are many popular relevant pages in English, it is not justifiable that Google does not retrieve such popular pages, even if they are in English, and the query is in Arabic.

The second evaluated metric on our query set was the monthly traffic on each query within the Arabic set of queries and their English translation. The Search-Based Keyword tool from Google was used to measure the monthly traffic on the selected queries and their English translations. It was found that, in 59% of the selected queries, Arabic users preferred to use English terms instead of Arabic ones in their queries. Extra analysis was performed on Google Insights for Search to go in depth in the traffic statistics on some Arabic queries and their translations. All Arabic queries in our query set as well as their English translations were analysed on the Google Insights for Search traffic estimator. Google Insights for Search statistics on Arabic queries and their translations gave us an indication about the query language preferences of Arabic users.

The results we got from analysing our queries on Google Insights for Search divided the Arabic queries into three clusters. The first is the popular English terms which have high traffic in Arabic countries. The second cluster is the popular Arabic terms which have high traffic in the Arab world, and finally the traditional Arabic queries which have high traffic. Analysis results of Google Insights for Search statistics show that, on average, about 77% of Arabic users prefer to use English terms when their queries contain terms from the first cluster (popular English terms) (i.e. Yahoo, Google, Facebook, YouTube). On the other hand when the query contains terms from the second cluster which have popular terms in Arabic, we found that 87% of the users prefer using these popular Arabic terms, *Maktoob* (مكتوب) and *Ajeeb* (عجيب), in their queries instead of their translated ones. Finally, when the query terms were from the third cluster, approximately 57% of users preferred using English terms in their queries (i.e. games (العاب) and software (برامج)). Further analysis of the above results was done to examine whether results set retrieved by Google takes into account the keyword popularity in English, such as the terms in the first cluster to retrieve the top results in English even if the query is in a different language (i.e. Arabic). The results of submitting a sample of our queries indicated that Google responses for particular queries may partially apply the statistics of traffic on keywords in their ranking techniques. As a result, for the majority of queries, the first result in Google response is in the language in which the query terms are submitted even if such query terms are not popular. The ultimate effects of not taking all traffic calculations on different languages into account may cause non-popular pages to appear first due to query language weights on Google ranking algorithm.

References

[1] Search by Keywords Tool, http://www.google.com/sktool/#, Google (visited on 10 August 2010).

[2] J. Ko, T. Mitamura and E. Nyberg, Language-independent probabilistic answer ranking for question answering. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, Association for Computational Linguistics, 2007), 784–91. query translation and text classification in a cross-language patent access system

[3] G. Bian and S. Teng, Integrating. *Proceeding of the 7th NTCIR Workshop Meeting* (NTCIR, Tokyo, 2008) 341–6.

[4] C. Huang, L. Chien and Y. Oyang, Query session based term suggestion for interactive web search. *Proceedings of the 10th WWW Conference* (10th WWW Conference, Hong Kong, 2001).

[5] B. Fonseca, P. Golgher, E. de Moura, B. Pôssas and N. Ziviani, Discovering search engine related queries using association rules, *Journal of Web Engineering* 2(4) (2004) 215–27.

[6] T. Xinhui, H. Tingting, L. Jing, C. Guang and Y. Zongkai, Chinese query expansion based on topic-relevant terms. *International Conference on Natural Language Processing and Knowledge Engineering*, 4(1) (NLP-KE, Beijing, 2008) 1–5.

[7] A. Al-Maskari, M. Sanderson and P. Clough, Arabic users' satisfaction with the online information as obtained from Google. *Proceedings of Sixth International Conference on Conceptions of Library and Information Science* (CoLIS, Borås, Sweden, 2007).

[8] Z. Chengyang, H. Yan, M. Rada and C. Hector, A natural language interface for crime-related spatial queries. *Proceedings of IEEE Intelligence and Security Informatics* (ISI, Dallas, TX, 2009).

[9] I. Janevski, K. Takasmanov and J. Pehcevski, NABU: a Macedonian web search portal. *In: Innovations in Information Technology*, IIT. Al Ain, United Arab Emirates (2008).

[10] P. Chew and A. Abdelali, The effects of language relatedness on multilingual information retrieval: a case study with Indo-European and Semitic languages. *Proceedings of the Workshop on Cross-Language Information Access* (Hyderabad, 2008).

[11] E. Airio, Who benefits from CLIR in web retrieval? *Emerald Group Publishing Limited* 64(5) (2008) 760–78.

[12] L. Dolamic and J. Savoy, Retrieval effectiveness of machine translated queries, *Journal of the American Society for Information Science and Technology* 61(11) (2010) 2266–73.

[13] P. Singh, S.C. Wight, O. Sercinoglu, D. Wilson, A. Boytsov and M. Raizada, Language preferences on websites and in google searches for human health and food information, *Journal of Medical Internet Research* 9(2) (2007) e18.

[14] L. Ballesteros and W. Croft, resolving ambiguity for cross-language retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, 1998) 64–71.

[15] K. Lee, K. Kageura and K. Choi, Implicit ambiguity resolution using incremental clustering in cross-language information retrieval, *Journal of Information Processing and Management* 40(1) (2004) 145–59.

[16] J. Gonzalo, F. Verdejo and I. Chugur, Using EuroWordNet in a concept-based approach to cross-language text retrieval, *Applied Artificial Intelligence* 13 (1999) 647–78.

[17] R. Hermes and F. Neto, Categorization-driven cross-language retrieval of medical information, *Journal of the American Society for Information Science and Technology* 57 (4) (2006) 501–10.

[18] L. Zhuhadar, O. Nasraoui, R. Wyatt and E.Romero, Multi-language Ontology-Based Search Engine, *The 3rd International Conference on Advances in Computer-Human Interactions,* (St. Maarten, The Netherlands, 2010), 13–18.

[19] M. Chau, J. Qin, Y.Zhou. C.Tseng. and H. Chen, SpidersRUs: Automated development of vertical search engines in different domains and languages, *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)* (Denver, Colorado USA, 2005).

[20] N. Efthimiadis, N. Malevris, N. Kousaridas, A. Lepeniotou and A. Loutas, An evaluation of how search engines respond to Greek language queries, *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (Waikoloa, HI, 2008), 136–9.

[21] H. Moukdad (2004) Lost In Cyberspace: How Do Search Engines Handle Arabic Queries? *Access to Information: Technologies, Skills, and Socio-Political Context*, (University of Manitoba, Winnipeg, Manitoba, June 3–5, 2004).

Appendix. Similarity between the collected queries for all Arabic characters during two periods of collection

Letter	Google Auto suggested list period (27/09/2009 to 7/10/2009)	Number of similar queries for the period (27/9/2010 to 7/10/2010)	Percentage of similar queries
ا	10	7	0.7
إ	10	5	0.5
أ	10	9	0.9
ى	10	9	0.9
ب	10	7	0.7
ت	10	6	0.6
ث	10	8	0.8
ج	10	7	0.7
ح	10	9	0.9
خ	10	6	0.6
د	10	7	0.7
ذ	10	7	0.7
ر	10	7	0.7
ز	10	6	0.6
س	10	8	0.8
ش	10	6	0.6
ص	10	5	0.5
ض	10	9	0.9
ط	10	5	0.5
ظ	10	6	0.6
ع	10	7	0.7
غ	10	7	0.7
ف	10	8	0.8
ق	10	8	0.8
ك	10	7	0.7
ل	10	9	0.9
م	10	7	0.7
ن	10	7	0.7
ه	10	6	0.6
ؤ	10	9	0.9
و	10	6	0.6
ي	10	9	0.9
ئ	10	7	0.7
		Average of similar queries	0.7

Source: Google auto-suggested list.