AN EVALUATION OF INTERNATIONAL WEB SEARCH ENGINES’ PERFORMANCES BASED ON ENGLISH, ARABIC, AND TURKISH LANGUAGES IN VARIOUS CATEGORIES

MOHAMAD YAHYA J. NAMK, YILTAN BİTİRİM

1Res. Asst., Department of Computer Engineering, Famagusta, T.R.N.C. 2Asst. Prof., Department of Computer Engineering, Famagusta, T.R.N.C. E-mail: 1yahya.namk@emu.edu.tr, 2yiltan.bitirim@emu.edu.tr

ABSTRACT

This study illustrates the performance evaluation of the three international Web search engines Google, Bing, and Yahoo based on the three languages English, Arabic, and Turkish in six various categories, i.e., “business”, “culture”, “health”, “sports”, “technology”, and “travel”. A total of 30 categorised queries (five per category) were determined and each of them was run in English, in Arabic, and in Turkish languages separately on each of the three Web search engines. At each run, every document retrieved in the first 10 was classified as “relevant” or “non-relevant” and, in each category, precision ratios were calculated for each of the five queries at every Web search engine and language pair. Then, they were used in the evaluation. Bing reached the greatest mean precision ratio (90%) for “health” category in English language. Yahoo reached the least mean precision ratio (32%) for “business” category in Turkish language. In all categories, generally, the Web search engines performed better performances in English document retrieved compared to Arabic document retrieved and Turkish document retrieved.

Keywords: Web Search Engine, English Language, Arabic Language, Turkish Language, Information Retrieval.

1. INTRODUCTION

In recent years, Web search engines have become the most important tools for the Web users. The importance of the Web search engines may come as a consequence of the huge number of pages on the Web. Gao et al. reported in 2013 that “search engines are the most important domain in Internet services in terms of the number of page views and daily visitors” [1]. Beside of these, Grosjean estimated in the study [2] that more than 50% of the world’s inhabitants use two or more than two languages (or dialects) in daily life; thus, it could be said that the Web users may use one or more than one languages in addition to their mother tongue. Furthermore, in the mid of 2014, there were almost 3 billion Internet users in the world which was around 40% of the total world’s inhabitants [3] and Internet users search for information in various categories according to their needs. Bringing all these together let us to have a study on evaluation of the performance of the three international Web search engines Google [4], Bing [5], and Yahoo [6] based on the three languages English, Arabic, and Turkish in six various categories, i.e., “business”, “culture”, “health”, “sports”, “technology”, and “travel”. The aims of this study are: to present information to the Web users who know one or more of these languages in order to have idea about the Web search engines and have advantage for better handling of the Web search engines; to motivate corresponding researchers in order to improve search engines; and to motivate search engine providers in order to improve their search engines.

Several related studies are as follows: Efthimiadis et al. [7] evaluated the five international Web search engines A9, AltaVista, Google, MSN, and Yahoo and the five Greek Web search engines Anazitisi, Ano-Kato, Phantis, Trinity, and Visto by using the navigational queries. These queries were homepage finding queries for known Greek organizations from the ten categories “government departments”, “universities”, “colleges”, “travel agencies”, “museums”, “media (TV, radio, newspapers)”, “transportation”, and “banks”. Furthermore, they evaluated the freshness of the
Web search engine indexes. One of their findings is that the international Web search engines performed better than the Greek search engines; Çakir et al. [8] evaluated the international image search engines Google, Yahoo, Ask, and MSN, for retrieving images based on various topics, “IT & Internet”, “Food & Beverage Brands”, “Broadcast Media”, “Automotive Manufacturers”, “Movies”, “Pharmaceutical & Medical Products”, and “Travel Destination & Accommodations”, with five English queries per topic. Mean precision ratios at 10, 20, 30, and 40 cut-off point values for each engine and topic pair were found and used in the evaluations; Goel and Yadav [9] presented an approach based on page level keywords for search engine evaluation. They used 40 educational queries and measured the performance of the international Web search engines Google, Yahoo, and Bing with their approach. Afterwards, they used the same queries and measured the performance of the same engines based on human ranking in terms of precision to verify their approach. Due to the similar results obtained from the two measurements, they declared that page level keywords could be good criteria to evaluate search engines; Moukdad [10] evaluated the performance of the international Web search engines Google, Altabhar, Ayna, and Morfix (the arabic module) based on their ability to retrieve documents containing morphologically related Arabic terms. The international Web search engines were limited in retrieving Arabic documents. Using the exact forms of Arabic words as search terms caused the loss of high number of documents.

The organization of this paper is as follows: the next section includes the methodology used in our study; section 3 demonstrates the experimental results; and the last section presents the conclusion of this study.

2. METHODOLOGY

In this study, we initially selected the highest ranked first three international Web search engines Google, Bing, and Yahoo as derived from the study of ebizMBA [11].

Three languages English, Arabic, and Turkish were selected. The reasons behind choosing these three languages were as follows: One of the authors knows all these three languages; in the Q-Success’ study [12], these three languages were recorded in the first 36 of “Usage-of-content-languages-for-websites” list which English language was ranked as the first (55.6%). Turkish language was ranked as eleventh (1.4%), and Arabic language was ranked as fourteenth (0.8%); and the numbers of English language speakers, Arabic language speakers, and Turkish language speakers were 335 million, 237 million, and 70.8 million, respectively [13].

Four popular international news providers CNN [14], euronews [15], Fox News [16], and The Huffington post [17] were selected. Then, six category topics, i.e., “business”, “culture”, “health”, “sports”, “technology”, and “travel”, were determined from these news providers. For each category, five specific queries were determined. Note that each query was especially considered to represent only its corresponding category that would be listed underneath. All queries (30 in total) in English, Arabic, and Turkish languages based on the six categories are given in table 1.

<table>
<thead>
<tr>
<th>Query #</th>
<th>Business</th>
<th>Culture</th>
<th>Health</th>
<th>Sports</th>
<th>Technology</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.</td>
<td>Brand licensing process</td>
<td>ancient greek gods</td>
<td>addiction treatment stages</td>
<td>billiards risks</td>
<td>cloud storage</td>
<td>first class flight features</td>
</tr>
<tr>
<td>A.</td>
<td>Free trade agreement</td>
<td>ethical symbols</td>
<td>breast cancer</td>
<td>boxing techniques</td>
<td>smart watch features</td>
<td>safari tours</td>
</tr>
<tr>
<td>T.</td>
<td>Export value</td>
<td>ethical symbols</td>
<td>breast cancer</td>
<td>boxing techniques</td>
<td>smart watch features</td>
<td>safari tours</td>
</tr>
<tr>
<td>E.</td>
<td>Investment risks</td>
<td>famous temples</td>
<td>obesity</td>
<td>derby matches</td>
<td>social media</td>
<td>travel tips</td>
</tr>
<tr>
<td>A.</td>
<td>Loan approval issues</td>
<td>famous temples</td>
<td>obesity</td>
<td>derby matches</td>
<td>social media</td>
<td>travel tips</td>
</tr>
<tr>
<td>T.</td>
<td>Tax changes</td>
<td>famous temples</td>
<td>obesity</td>
<td>derby matches</td>
<td>social media</td>
<td>travel tips</td>
</tr>
<tr>
<td>E.</td>
<td>Professional salesmanship</td>
<td>traditional clothes</td>
<td>vanity</td>
<td>swimming styles</td>
<td>yosemite operating system</td>
<td>yatch voyages</td>
</tr>
<tr>
<td>A.</td>
<td>Marketing</td>
<td>traditional clothes</td>
<td>vanity</td>
<td>swimming styles</td>
<td>yosemite operating system</td>
<td>yatch voyages</td>
</tr>
<tr>
<td>T.</td>
<td>Professional services</td>
<td>traditional clothes</td>
<td>vanity</td>
<td>swimming styles</td>
<td>yosemite operating system</td>
<td>yatch voyages</td>
</tr>
</tbody>
</table>

Notes: “E.” stands for English; “A.” stands for Arabic; “T.” stands for Turkish.
Before run the queries, some preferences of the three Web search engines were changed: (*) The language setting of the Web search engines was changed based on the query language to retrieve the search results in the same language of the query; (*) SafeSearch feature were turned off to reach the search results without any filters; (*) The location setting was specified to retrieve the search results from all over the world.

The first query was run in English, in Arabic, and in Turkish languages one by one on each of the three Web search engines. At each run, the first 10 documents retrieved were evaluated. The same were done for the rest of the queries separately and, as a result, totally 270 (30 queries × 3 languages × 3 Web search engines) runs and retrieval output evaluations were done. During the evaluation of each retrieval output, every document retrieved in the first 10 was categorised as “relevant” or “non-relevant” based on the followings: (*) If the text content of a document was related to subject of the query, it was categorised as “relevant”; (*) if a document was repeated at least one time with different URL address – the URL addresses appeared in the address bars were considered after the documents were displayed –, all the documents (repeated-document and repeat-document(s)) were categorised as “relevant” or categorised as “non-relevant” based on the relation of the text content with the query subject (note that “SD” code will be used to represent a repeat-document in a such case); (*) if a document was repeated at least one time with the same URL address – the URL addresses appeared in the address bars were considered after the documents were displayed –, the repeated-document was categorised as “relevant” or “non-relevant” based on the relation of the text content with the query subject, while the repeat-document(s) was directly categorised as “non-relevant” (note that “RD” code will be used to represent a repeat-document in a such case); (*) the documents were categorised as “non-relevant” based on the followings: (+) If a document’s text content was not related to subject of the query; (+) if a document’s text content was in different language other than the query language; (+) if a document included images or videos without any text description related to the query; (+) if a document required registration for accessing the content; (+) if a document was appeared with the message “Access Denied” or the message “509 Bandwidth limit exceeded, the server is temporarily unable to serve your request due to the site owner reaching his/her bandwidth limit”; (+) if a document was not accessible and one of the following messages was displayed: “This website is not available”; “No data received”; “URL unavailable, please contact the administration team if you think it’s an error”; “404 error”. Note that this situation was considered as a dead link situation and from now on, “DD” code will be used to represent such situation.

Precision ratio is the number of relevant documents retrieved divided by the total number of documents retrieved which all multiplied by 100 [18]. In each category, precision ratios were calculated for each of five queries at every Web search engine and language pair. For each Web search engine, mean precision ratios were calculated for each category and language pair by using the precision ratios of the five queries in the corresponding category. Additionally, average mean precision ratios of the six categories for each Web search engine and language pair were calculated.

3. EXPERIMENTAL RESULTS

When considering all query runs, 2700 documents were retrieved in total. 22 of them were DDs, 20 of them were RDs, and 33 of them were SDs. Excluding DDs, 2678 documents were examined and 1840 of them were categorised as “relevant”.

For each Web search engine, table 2 shows the total number of DDs, RDs, and SDs retrieved from 5 queries per a category for each language. For all three languages, Google retrieved the lowest total number of DDs (5), RDs (2), and the SDs (5) among the engines. Contrary to Google, Yahoo retrieved the highest total number of DDs (9) and RDs (11) for all three languages, while both Yahoo and Bing retrieved the same total number of SDs (14) for all three languages. Every Web search engine retrieved at least total one DD for all queries per language. Google retrieved total zero number of RDs and SDs for all queries in English and Turkish languages, respectively. On the other hand, Google retrieved total one RD for all queries in both Arabic and Turkish languages and at least total two SDs for all queries in both English and Arabic languages. Both Bing and Yahoo retrieved at least total one RD and at least total one SD for all queries in each language. When all Web search engines are considered, the lowest category-based total numbers of DDs, RDs, and SDs for all queries
in all three languages were 1 in “culture”, “health”, and “technology” categories, 1 in “business”, and “travel” categories, and 2 in “technology” category, respectively, while the highest category-based total numbers of DDs, RDs, and SDs for all queries in all three languages were 9 in “business”, 7 in “technology”, and 11 in “travel”, respectively.

As table 3 shows, in all categories, Google retrieved at least one “relevant” document for every query and language pair. Bing retrieved zero “relevant” documents only for query 5 of the “travel” category in Arabic language. Yahoo retrieved zero “relevant” documents for queries 5, 2, and 4 of “business”, “culture”, and “travel” categories, respectively in Turkish language, and for queries 2 and 5 of “sports” and “travel” categories, respectively in Arabic language. Google retrieved 10 “relevant” documents for one or more queries in every category, whereas Bing and Yahoo have the same except in “culture” and “technology” categories, respectively. Only Yahoo retrieved the same number of “relevant” documents in all languages for the fourth query of “sports” category and fifth query of “technology” category. When each language-based-totals are considered, it is seen that Bing has the highest total number of “relevant” documents (45) in “health” category for English language. In contrast, Yahoo has the lowest total number of “relevant” documents (16) in “business” category for Turkish language.

Figure 1 demonstrates mean precision ratios of Google for every category and language pair. In all categories except “travel” and “health”, Google presented the best performance in English language. In “travel” category, it presented the best performance (86%) in Turkish language, while in “health” category, it displayed the best performance (84%) in both English and Arabic languages. Besides, in “health” and “technology” categories, Google presented the lowest performance in Turkish language, while for the rest of the categories, it presented the lowest performance in Arabic language. Google’s mean precision ratios range for all languages was between 52% and 88%. The best mean precision ratio (88%) was shown in “culture” category in English language and the lowest mean precision ratio (52%) was shown in
both “business” and “travel” categories in Arabic language.

Figure 1: Mean Precision Ratios of Google Web Search Engine

Figure 2 clarifies mean precision ratios of Bing for every category and language pair. In all categories, Bing showed the best performance in English language among all languages. In “culture” and “sports” categories, Bing showed the lowest performance in Arabic language, while for the rest of the categories, it showed the lowest performance in Turkish language. Bing’s mean precision ratios range for all languages was between 42% and 90%. The best mean precision ratio (90%) was shown in “health” category in English language and the lowest mean precision ratio (42%) was shown in “culture” category in Arabic language.

Figure 2: Mean Precision Ratios of Bing Web Search Engine

Figure 3 illustrates mean precision ratios of Yahoo for every category and language pair. With the exception of “health” category, in all categories, Yahoo displayed the best performance in English language. In “health” category, the best performance (82%) was displayed in both English and Arabic languages. Excluding “sports” category, in all categories, Yahoo has the lowest performance in Turkish language. In “sports” category, the lowest performance was displayed in Arabic language. Yahoo’s mean precision ratios range for all languages was between 32% and 86%. The best mean precision ratio (86%) was shown in “sports” category in English language and the lowest mean precision ratio (32%) was shown in “business” category in Turkish language.

Figure 3: Mean Precision Ratios of Yahoo Web Search Engine

Figure 4 demonstrates mean precision ratios of Google, Bing, and Yahoo in English language for all categories. In all categories except “culture” and “travel”, Bing presented the best performance. In both “culture” and “travel” categories, Google presented the best performance. In both “business” and “sports” categories, Google presented the worst performance. In “health”, “technology”, and “travel” categories, Yahoo showed the worst performance. In “culture” category, Bing has the worst performance. The mean precision ratios range for English language was between 64% (displayed by Yahoo in “travel” category) and 90% (displayed by Bing in “health” category).

Figure 4: Mean Precision Ratios for English Language

Figure 5 clarifies mean precision ratios of Google, Bing, and Yahoo in Arabic language for all categories. In “business” category, the best performance was shown by both Bing and Yahoo, in “sports” category, by both Google and Bing, and in “culture” category, by Google, while in the rest of the categories, Bing showed the best performance. In “culture” and “health” categories, Bing and Yahoo displayed the worst performance, respectively. In both “technology” and “travel” categories, Google and Yahoo have shared the worst performance. The mean precision ratios range for Arabic language was between 42% (displayed by Bing in “culture” category and displayed by...
Yahoo in “sports” category) and 88% (displayed by Bing in “health” category).

Figure 5: Mean Precision Ratios for Arabic Language

Figure 6 illustrates mean precision ratios of Google, Bing, and Yahoo in Turkish language for all categories. In all categories except “sports” and “technology”, Google displayed the best performance. In both “sports” and “technology” categories, Bing displayed the best performance. In all categories except “technology” which Google has the worst at it, Yahoo has the worst performance. The mean precision ratios range for Turkish language was between 32% (displayed by Yahoo in “business” category) and 86% (displayed by Google in “travel” category).

Figure 6: Mean Precision Ratios for Turkish Language

4. CONCLUSION

The performance evaluation of the three international Web search engines Google, Bing, and Yahoo in the three languages English, Arabic, and Turkish in “business”, “culture”, “health”, “sports”, “technology”, and “travel” categories were studied.

The highest category-based total number of DDs was in “business” category, RDs in “technology” category, and SDs in “travel” category, while the lowest category-based total number of DDs was in “culture”, “health”, and “technology” categories, RDs was in “business” and “travel” categories, and SDs in “technology” category.

The highest total number of “relevant” documents was retrieved by Bing in “health” category for English language. In contrast, the lowest total number of “relevant” documents was retrieved by Yahoo in “business” category for Turkish language.

In “health” category, both Google and Yahoo displayed their best performances in both English and Arabic languages, while Bing showed its best performance only in English language. In “travel” category, Google showed its best performance in Turkish language, while both Bing and Yahoo showed their best performances in English language. In the rest of the categories, all three Web search engines showed their best performances in English language.

When English language is considered, Google presented the highest performance in both “culture” and “travel” categories, while Bing presented the highest performance in the rest of the categories. For Arabic language: both Yahoo and Google shared the highest performance in “business” category; both Bing and Google shared the highest performance in “sports” category; Google presented the highest performance in “ culture” category; and Bing presented the highest performance in the rest
of the categories. When Turkish language is considered, Bing presented the highest performance in both “sports” and “technology” categories, while Google presented the highest performance in the rest of the categories.

The greatest mean precision ratio (90%) reached by Bing for “health” category in English language. On the other hand, the least mean precision ratio (32%) reached by Yahoo for “business” category in Turkish language.

For all categories, generally, the Web search engines performed better performances in retrieving English documents. Nevertheless, we believe that international Web search engines still require amelioration in order to retrieve more “relevant” documents (especially in Arabic and Turkish languages) in various categories for increasing the user satisfaction on retrieval effectiveness.

REFERENCES