Asian Journal of Information Technology 14 (3): 105-110, 2015
ISSN: 1682-3915

AEHURA: A New Ranking Algorithm for Arabic Web Search Engines

Safaa I. Hajeer, Rasha M. Ismail, Nagwa L. Badr and M.F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

Abstract: Most search engine systems mainly focusing on developing Western languages such as English so these search engine systems have a high performance on these languages but they don't have the same performance when they are used for Eastern languages such as Arabic. Furthermore, Arabic is a highly inflected language that has a complex morphological structure, so search engines always have a challenge to find the best ranked list to the user's query from those huge numbers of pages. A lot of search results that correspond to a user's query are not relevant to the user's need. Most of the page ranking algorithms use link-based ranking (web structure) or content-based ranking to calculate the relevancy of the information to the user's need but those ranking algorithms might not be enough to provide a good ranked list for the Arabic search. So in this study, we proposed an efficient Arabic information retrieval system using a new Arabic hybrid usage-based ranking algorithm called AEHURA. The objective of this algorithm is to overcome the drawbacks of the ranking algorithms and improve the efficiency of Arabic web searching.

Key words: Information Retrieval (IR), tokenization, search engine, indexing, Arabic language, usage-based ranking, content-based ranking, link-based ranking

INTRODUCTION

Over the years search engines play critical roles in our lives all around the world. The importance of web search engines may come as a consequence of the huge number of pages on the web. These huge numbers of pages are a variety in contents, most of them in English language but the rest are in several different languages like German, Hindi, Russian, Arabic, Turkish, etc.

Recently, the changing in the requirements of every day life, reflects on the growing number of internet users around the globe, Information Retrieval (IR) has become of great importance as an essential tool for all tasks of searching on the web. Thus, the number of Arab internet users has increased recursively over the years. Arabic is one of the six official languages of the United Nations and the mother tongue of >360 million people that are spread over 22 countries (Dilekh and Behloul, 2012). Arabic is a highly inflected language and has a complex morphological structure; the Arabic alphabet consists of 28 letters and can be extended to ninety by additional shapes, marks and vowels. Each letter can appear in up to four different shapes depending on whether it occurs at the beginning, middle or at the end of a word or alone (Meftouh *et al*., 2010). So, it is very problematic to build an information retrieval system on the Arabic language due the specific morphological and structural changes in the language: irregular and inflected derived forms, various spellings of certain words, various writing of certain combination characters and the short and long vowels. As a result, the performance of search engines varies with the language used and depends on the nature and the complexity of the language in which the request of research is formulated. The operation of an engine is mainly based on an automatic treatment of the natural language. These treatments differ from one language to another and may depend on the particular characteristics of this language. Most of the available engines which are primarily developed for the Western languages such as English are increasingly powerful in these languages. In addition, these performances are less in the case of the Arabic language, probably because of specificities morphological and structural characteristics of Arabic compared to the Western languages. Indeed, few studies have focused on studying its performance in the Arabic language. Also, most existing web search engines often calculate the relevancy of web pages for a given query by counting the search keywords contained in the web pages, this approach is called content-based ranking algorithms that use the words in each document to determine its ranking. This approach works well when users' queries are clear and specific. However, in the real world, web search queries are often short (<3 words) and ambiguous and web pages contain a lot of diverse and noisy information (Jiang *et al*., 2005). These will very likely lead to the deterioration in the performance of web search engines, due to the gap between query space

Corresponding Author: Safaa I. Hajeer, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

and document space. Another approach, link-based ranking algorithms assign scores to web pages based on the number and quality of hyperlinks between pages. Links that point to a particular page or endorse a page can help to improve the link-based rankings. Finally, usage-based ranking algorithms score documents by how often they are viewed by internet users. For usage-based ranking, there is limited work to utilize the usage data in the web information retrieval systems, especially in the ranking algorithm. For some systems (Rodriguez-Mula *et al*., 1998) that do use the usage data in ranking, they determine the relevance of a web page by its selection frequency. This measurement is not that accurate to indicate the real relevance. The time spent on reading the page, the operation of saving, printing the page or adding the page to the bookmark and the action of following the links in the page are all good indicators, perhaps better than the simple selection frequency, so it is worth further exploration on how to apply this kind of actual user behavior to the ranking mechanism. For these reasons, this research has provided a Hybrid Ranking algorithm for Arabic search engines called AEHURA (Arabic Efficient Hybrid Usage-based Ranking Algorithm), this algorithm basically utilizes the usage data with Arabic web pages. AEHURA is proposed to improve the ranked list provided from search engines that are based only on the content base rankings. This improvement will have a direct effect on the effectiveness and the performance of Arabic information retrieval systems and Arabic web search engines too.

LITERATURE REVIEW

Many efforts are done in order to have algorithms for improving search engine results. Each research is always trying to prove the effectiveness on search engines all around the globe.

Kritikopoulos studied a method for evaluating the quality of ranking algorithms. Success index takes into account a user's click through data and the result shows their method is better than explicit judgment.

A comparison study was proposed by Liu *et al*. (2010) between three methods of ranking in the usage field. Those methods are PageRank, Weighted PageRank and HITS. All of those methods focus on the structure of the page and the result of this comparison shows that HITS is the best.

Elraouf *et al*. (2010) presented an enhanced ranking algorithm by combining both the content-based algorithm and link-based algorithm with the focus on an Arabic search engine, this study idea is based on combining both the count of words related to a query in the page and the count of words related to a query in out link pages of that page to calculate its rank. The results show their proposed algorithm for ranking is better than the Google search engine ranking by 3.2 times, its also better than the Yahoo search engine ranking by 4.3 times.

Jain and Purohit (2011) presented a method based on a combination of click-through of pages by the users (event) and the summarization of documents. They used the advantage of implicit modelling as this effectively improves the user's model without any extra effort of the user as a result, implicit feedback information improves the user modelling process. Another study was presented by Iyakutti (2011). This study provided a new model to find a user's preferences from click-through behavior and used the exposed preferences to adapt the search engine's ranking function for improving the search service. In this proposed model, the combination of viewed and stored document summaries is used. The results show that this combining improved the reliability of the ranked list more than ever before.

Mukherjee presented a method to discover web knowledge for presenting web users with a more personalized web content. Their method collected usage data from different users and then found the similarities between all pairs of users. Experimental results generate correct suggestions that retrieve relevant documents to the user.

Tuteja (2013)'s study was based on user behaviors in order to enhance the weighted PageRank algorithm by considering a term Visits of Links (VOL) done by the end of 2013. This research idea was presented as modifying the standard Weighted PageRank algorithm by incorporating visits of links. The result shows that adding the number of Visits of Links (VOL) to calculate the values of page rank proves that relevant results are retrieved first. In this way, it may help users to get the relevant information much quicker.

In 2014, a research done by Hajjar, this research was interested in studying the performance of search engines which were the most famous between 2006 and 2010, Google, Yahoo, Copernic, Bing ask, AOL Search and MSN/Live, based on a corpus of a thousand Arabic documents. The results showed that the search engine Google in its local version can extract only those documents that contain exactly the query word. The search engines: Windows Search, X1, Copernic Search, AOL search in their local versions and gave the same results under the condi-tions of their experiments.

A new approach is introduced by Patil and Keole (2012) to re-rank the search results list based on the

contents and user's interests rather than keyword and page ranking provided by search engines. When the user visits the web page out of this re-rank list, the query, url and the contents extracted from the web page are stored in the server log. When the next time the user enters a query, the scores are awarded to each result link based on the data in the server log which indirectly incurs the user's interest. Hajeer proposed an English Hybrid Usage-based Ranking Algorithm called EHURA. EHURA was applied to 1033 English Corpus to measure its performance. The result shows EHURA improves the precision over the content based algorithm by about 15% while realizing approximately the same recall percentage. From previous, few researches considered the usage-based ranking in English and very few of them are concerned with Arabic usage-based ranking algorithms. Indeed, few studies have focused on studying the search engines performance in the Arabic language at all. On the other hand, most researches are based on the pages selection frequency. This might be an incorrect indicator; the reasons might be inadvertent human mistakes, misleading titles of web pages or the returned summaries not representing the real content. As a conclusion, ranking algorithms still have some draw-backs in the ranked list provided by some search engines. So, we propose a Hybrid Ranking algorithm to utilize the usage data and apply it to a Arabic search engine as Arabic is a highly inflected language and has a complex morphological structure which makes information retrieval on Arabic texts a challenge. This Hybrid Ranking algorithm is based on content-based ranking which is the more accurate indicator instead of the link-based ranking. The algorithm is based on the content of the pages ranked list and in addition to other usage factors which are:

- Frequency of visit that determines the relevance of a web page by its selection frequency
- Time Spent shows how long users spend on a page after removing the download time of the page
- Click Event is the click history of a page to assign a quantitative weight to each page for a user

THE PROPOSED SYSTEM ARCHITECTURE

This study discusses the proposed Arabic search engine system; the basic idea of the system is based on the Arabic Efficient Hybrid Usage-based Ranking Algorithm (AEHURA). This Hybrid Ranking algorithm is to improve the ranked list provided from search engines that are based only on content base ranking. The system architecture is shown in Fig. 1. The proposed system consists of 5 main modules:

Module 1: The preprocessing steps include: detecting Arabic texts in different encodings and converting texts to a common encoding; handling some of the ortho-graphic features; tokenization and data cleaning by handling useless words (stop-words) far away from the useful keywords in searching; performing stemming and handling lexical and spelling variations and finally performing the best indexing for Arabic's terms. The stages of preprocessing of our methodology as the following:

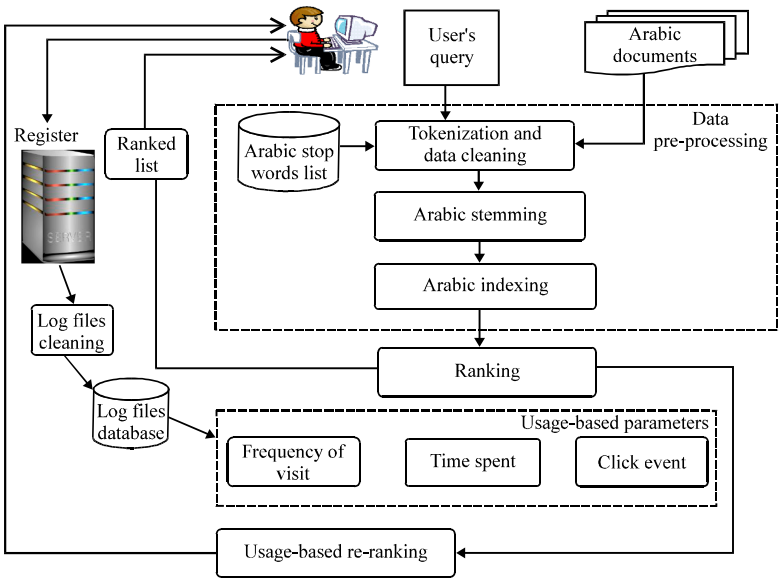

Fig. 1: The proposed system architecture

Stage 1 (Tokenization and data cleaning): This stage takes the Arabic text documents and the user's query and split their Arabic text into a stream of words using the parsing technique and breaking of white spaces then it keeps the words in a list called a Word's list, after that the data cleaning removes the useless words from this word's list that occur in 80% of the document (Maurya *et al*., 2013), these useless words are stored in a stop words database as appears in the figure. The database has 1459 stop words with a size of 10 kB.

Stage 2 (Arabic stemming): The stemming process reduces the size of the docu-ment representations by 20-50% compared to full word representations, according to Van Rijsbergen (Sembok *et al*., 2011). The Hybrid Affix Removal algorithm is applied in this module, this Stemmer algorithm is between root-based and light stemmers. It removes the suffixes and prefixes of Arabic words and in addition it returns some words to their basic roots as by Arafat and Saad (2008).

Stage 3 (Arabic indexing): Arabic indexing is a process for describing or classifying an Arabic document by index terms. These index terms are grouped in an indexer and the Arabic stemmer services this stage by improving the group of these keywords in the indexer.

Module 2 (Ranking): In this module, the user's query is matched with the index terms to get the relevant documents to the query. Documents are then ranked using ranking algorithms according to the most relevant to the user's query. The ranking algorithm here is the Content-based Ranking algorithm which simply tries to find the similarity between the content of the documents and the query. We applied here the cosine similarity measure, this selection is based on studies represented by Hajeer (2012).

Module 3 (Log files cleaning): The log file contains lots of irrelevant entries which need to be removed. To enhance the efficiency of usage-based retrieval, any noise should be removed (such as page moved permanently), files that do not exist, server internal errors, service temporarily unavailable, etc.) before retrieving the usage data.

Module 4 (Usage-based parameters): In this module, the system calculates three usage-based parameters as follows: frequency of visit that determine the relevance of a web page by its selection frequency:

$$FW = \frac{\text{No. of visit on a page (u)}}{\text{Total number of visit on all page}} \times PR(u) \quad (1)$$

Where:
FW = Frequency Weight
PR(u) = The page rank of a page u

Time spent shows how long users spend on a page after removing the download time of the page:

$$TW = \frac{\text{Time spent on a page (u)-Download time (u)}}{\text{Max (Time spent on a page (u)-Download time (u))}} \quad (2)$$

where, TW is time spent weight.

$$\text{Download time (u)} = \frac{\text{Size of a page (u)}}{\text{Transfer rate for page (u)}} \quad (3)$$

Click event is the click history of a page to assign a quantitative weight to each page for a user:

$$CEW = \begin{cases} 0.5 & \text{if an event is done} \\ 0 & \text{otherwise} \end{cases}$$

where, CEW is Click Event Weight.

Module 5 (Usage-based re-ranking): This module is a re-ranking process based on the combination (i.e., the summation) of the above parameters (frequency of visit, time spent and click event) with the content-based ranking results provide for our EHURA algorithm which is trying to improve the ranking result from Arabic Web search engines.

PERFORMANCE STUDIES

In order to study the performance of the proposed system, we used different evaluation measures. These measures are discussed in evaluation IR System. Then, the data sets used and the experimental results are shown in experimental results.

Evaluation IR System: In order to measure the performance of our IR System, we evaluate our proposed EHURA algorithm by measuring the performance of it then comparing its results with the content-based ranking algorithm. The performance measured by the recall and precision measurements and other measures are represented in the following Eq. 5 and 6:

$$\text{Precision} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retrived documents}\}|}{|\{\text{Retrived documents}\}|} \quad (5)$$

$$\text{Recall} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retrived documents}\}|}{|\{\text{Relevant documents}\}|} \quad (6)$$

Fall-out is the proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

$$\text{Fall-out} = \frac{|\{\text{Non-relevant documents}\} \cap \{\text{Retrived documents}\}|}{|\{\text{Non-relevant documents}\}|} \quad (7)$$

$$\text{F-measure} = \frac{|2 \times \text{Precision} \times \text{Recall}|}{|\text{Precision} + \text{Recall}|} \quad (8)$$

where, F-measure is the weighted harmonic mean of precision and recall:

$$\text{AveP} = \frac{\sum_{i=1}^{N_0} P_i(r)}{N_q} \quad (9)$$

Where:

AveP = Average precision at recall level r

$P_i(r)$ = The precision at recall level r for the ith query

N_q = The number of queries used

Experimental results: For testing the proposed system, it was applied on the Ain Shams Arabic corpus. This corpus belongs to the modern standard Arabic type; it contains 242 documents with different sizes, the system was tested with 20 queries in order to evaluate the IR System performance.

The system was tested using IR evaluation measurements which was mentioned in the evaluation section. Figure 2 shows the precision and recall results for each query for the content-based algorithm in comparison with the hybrid algorithm (AEHURA). It's clear that AEHURA reaches a better result than the content-based one. The average precision of our new approach (AEHURA) reached 98% while the precision of the Content-based Ranking algorithm is 88%, the results are shown in Table 1. So, our proposed AEHURA algorithm improves the precision over the Content-based Ranking algorithm by about 10% while it also improves the recall percentage by 7%.

The proportion of non-relevant documents retrieved (Fall-out) from the system using the content-based algorithm reached 29% while our proposed EHURA algorithm reached 22%.

Figure 3 shows the F-measure for using the content-based algorithm and the EHURA and it's clear from the Fig. 3 that the EHURA algorithm improved the F-measure over the content-based algorithm by 8%.

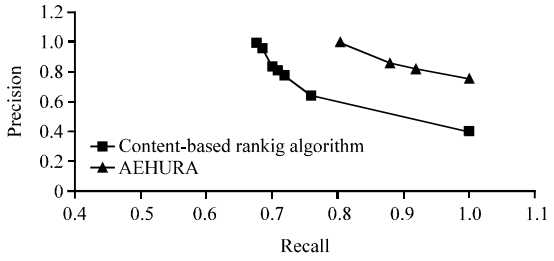

Fig. 2: Precision and recall for ranking against the Arabic Ain Sham's corpus 20 queries

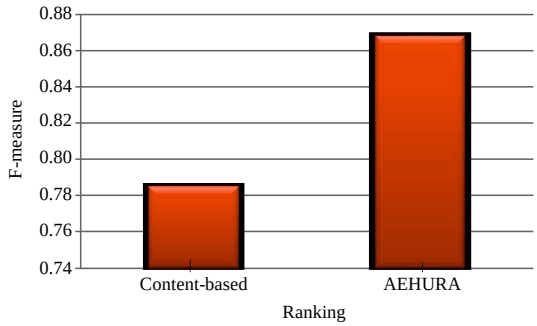

Fig. 3: F-measure for ranking against the Arabic Ain Sham's corpus

Table 1: Evaluation summary for the proposed system

Summary	Precision	Recall	Fall-out	F-measure
Content-based ranking	0.8753	0.7125	0.2875	0.7856
EHURA	0.9802	0.7800	0.2200	0.8687

CONCLUSION

Arabic is a highly inflected language that has a complex morphological structure. In order to develop an Arabic search engine is a challenge. Many researches are done to improve the ranking algorithm for search engine systems are based on western lan-guages like English. Thus, these search engine systems have a high performance on these languages but they don't have the same performance when they are used for Eastern languages such as Arabic. In addition, few researches have been done on Arabic search engines. In this study, we proposed an efficient Arabic information retrieval system using a new Arabic Hybrid Usage-based Ranking algorithm called AEHURA. The objective of this algorithm is to overcome the drawbacks of the ranking algorithms and improve the efficiency of Arabic web searching.

The system was applied to the Ain Shams Arabic corpus for testing and evaluation. The results show that the AEHURA algorithm improves the performance of the information retrieval system in respect to the recall and precision measures. It also improves the precision over the content-based ranking algorithm by about 10% while improving the recall percentage by 7%.

REFERENCES

Arafat, S. and S. Saad, 2008. An affix removal stemming algorithm for Arabic language. Int. J. Intell. Comput. Inf. Syst., 8: 141-153.

Dilekh, T. and A. Behloul, 2012. Implementation of a new hybrid method for stemming of Arabic text. Int. J. Comput. Appl., 46: 14-19.

Elraouf, E.A., N.L. Badr and M.F. Tolba, 2010. An efficient ranking module for an Arabic search engine. IJCSNS, 10: 1-218.

Hajeer, S.I., 2012. Comparison on the effectiveness of different statistical similarity measures. Int. J. Comput. Appl., 53: 14-19.

Iyakutti, K., 2011. Improving the information retrieval system through effective evaluation of web page in client side analysis. Int. J. Comput. Appl., 15: 35-39.

Jain, R. and D.G. Purohit, 2011. Page ranking algorithms for web mining. Int. J. Comput. Appl., 13: 0975-8887.

Jiang, X.M., W.G. Song and H.J. Zeng, 2005. Applying Associative Relationship on the Clickthrough Data to Improve Web Search. In: Advances in Information Retrieval. David, E. and M. Juan (Eds.). Springer, Berlin Heidelberg, Germany, ISBN: 978-3-540-25295-5, pp: 475-486.

Liu, Y., T.Y. Liu, B. Gao, Z. Ma and H. Li, 2010. A framework to compute page importance based on user behaviors. Inf. Retrieval, 13: 22-45.

Maurya V., P. Pandey and L.S. Maurya, 2013. Effective information retrieval system. Int. J. Emerging Technol. Adv. Eng., 3: 787-792.

Meftouh, K., M. Tayeb Laskri and K. Smaili, 2010. Modeling Arabic language using statistical methods. Arabian J. Sci. Eng., 35: 1-69.

Patil, S.C. and R.R. Keole, 2012. Content and usage based ranking for enhancing search result delivery. Int. J. Sci. Res., Vol. 3.

Rodriguez-Mula, G., H. Garcia-Molina and A. Paepcke, 1998. Collaborative value filtering on the web. Comput. Networks, ISDN Syst., 30: 736-738.

Sembok, T., B. Abu Ata and Z. Bakar, 2011. A rule and template based stemming algorithm for Arabic language. Int. J. Mathemtical Models Methods Appl. Sci., 5: 974-981.

Tuteja, S., 2013. Enhancement in weighted pagerank algorithm using VOL. IOSR J. Comput. Eng., Vol. 14.