International Journal of Soft Computing 12 (4): 280-286, 2017
ISSN: 1816-9503

A Hybrid Usage-Based Ranking for Enhancing Arabic 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: There are billions of web pages available on the internet. 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 needs. 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 be not 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 Hybrid Usage-Based Ranking algorithm called EHURA. The objective of this algorithm is to overcome the drawbacks of the ranking algorithms and improve the efficiency of web searching. EHURA was applied to 242 Arabic corpus to measure its performance. The result shows our proposed EHURA algorithm improves the precision over the Content-Based Ranking algorithm representation as well as the recall is affected too in this improvement.

Key words: Information Retrieval (IR), usage-based ranking, content-based ranking, linked-based ranking, weighted page rank

INTRODUCTION

The amount of information in the world is increasing exponentially through the years. Searching within this huge amount of information becomes a critical behavior of our life. Millions of users interact with search engines daily around the globe; >360 of them are Arab ones (Dilekh and Behloul, 2012).

Recently, due to the growing number of internet users around the world, Information Retrieval (IR) has become of great importance as an essential tool for all tasks of searching on the web. The number of Arab internet users has increased recursively over the years because of the changes in the requirements of the life. Relatively fewer Arabic search engines are currently available, despite the enormous efforts to satisfy the needs of the growing number of Arabic internet users. Moreover, Arabic is a highly inflected language and has a complex morphological structure which makes information retrieval on Arabic texts a challenge (Dilekh and Behloul, 2012).

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 search engines 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 (Hajjar *et al*., 2010). Indeed, few studies have focused on studying its performance in the Arabic language search engine and its algorithms.

Really, each search engine's algorithm is unique so a top ranking on Yahoo! does not guarantee a prominent ranking on Google and vice versa. To make things more complicated, the algorithms used by search engines are not only closely guarded secrets, they are also constantly undergoing modification and revision. This means that the criteria to best optimize a site with must be surmised through observation as well as trial and error and not just once but continuously (Anonymous, 2015).

Also, most of 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 user's queries are clear and specific. However, in the real world, web search queries are often short (<3 words) and ambiguous (Jiang *et al*., 2005) and web pages contain a lot of diverse and noisy information. These will very likely

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

lead to the deterioration in the performance of web search engines due to the gap between query space 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 (Ding *et al*., 2002; 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. The objective of the study is to provide a Hybrid Ranking algorithm to utilize the usage data called EHURA (Efficient Hybrid Usage-Based Ranking Algorithm). This ranking algorithm is proposed to improve the ranked list provided from search engines that are based only on content base rankings. This improvement will have a direct effect on the effectiveness and the performance of information retrieval systems and web search engines.

Literature view: In the context of information retrieval, ranking algorithms have become the most researched area of information retrieval. Many researchers have developed algorithms for improving search engine results. Each research proposed a methodology and measurements to test the performance and compute the accuracy of its algorithm.

Kritikopoulos *et al*. (2007) was studied method in 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 page rank, weighted page rank and HITS. All of those methods focus on the structure of the page. The result of this comparison shows that HITS is the best.

By Jain and Purohit (2011), this research 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 is effectively improving the user model without any extra effort of the user as result implicit feedback information improves the user modelling process.

Another study was presented by Rekha. 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 than ever was (Lebbos *et al*., 2014).

Mukherjee *et al*. (2012) presented a method to discover web knowledge for presenting web users with 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 (Mukherjee *et al*., 2012).

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 new approach is introduced in (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.

A research done by Hajjar *et al*. (2010) this research was focused in studying the performance of search engines 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 conditions of their experiments.

In 2015, Hajeer *et al*. (2015) 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 it is rare of them are concerned in Arabic Usage-Based Ranking algorithms. 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 drawbacks in the ranked list provided by some search engines. So, we decide to develop a Hybrid Ranking algorithm to utilize the usage data and apply it to an Arabic search engine 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 in addition to other usage factors which are:

- Frequency of visit that determine 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-through is the click history of a page to assign a quantitative weight to each page for a user

MATERIALS AND METHODS

The system architecture: This study discusses the proposed Arabic search engine system; the basic idea of the system is based on the Efficient Hybrid Usage-Based Ranking Algorithm (EHURA). 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 (tokenization): This module is parsing the content of text documents and breaking the stream of text into words, then keeping the words in a list called a word's list. This module is parsing the content of text documents and breaking a stream of text of them into words then keeping the words in a list called a word's list, i.e., documents in fact are a sequence of words that have ideas; computers don't understand the structure of a natural language document and can't automatically recognize words and sentences unlike humans. To a computer, a document is only a sequence of bytes. Computers don't know that a space character separates words in a document; instead, humans must program the computer to identify what constitutes an individual or

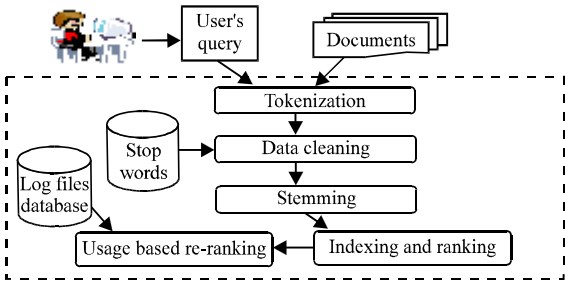

Fig. 1: The proposed system architecture

distinct word referred to as a token. This operation is called tokenization. The user's query on the system acts with it like a document.

Module 2 (data cleaning): The data cleaning module removes useless words from the word's list; these useless words may punctuation marks, prepositions, pronouns, conjunctions and auxiliary verbs which referred to as stop words and often removed. These stop words are deemed irrelevant for searching purposes because they occur frequently in the language for which the indexing engine has been tuned, these words are dropped at indexing time and then ignored at search time. For example have, did in, of, get, etc. are meaningless, the useless words are stored in a stop words database as appears in Fig. 1. The database has 1459 stop words with a size 10 KB.

Module 3 (stemming): The Hybrid Affix Removal algorithm is applied in this module this Hybrid Affix Removal algorithm is explained in detail in next study.

Module 4 (indexing and ranking): Indexing is a process for describing or classifying a document by index terms; index terms are the keywords that have their own meaning (i.e. which usually has the semantics of the noun).

These index terms are grouped in an indexer and a stemmer services this stage by improving the group of these keywords in the indexer. Then, the user's query is matched with the index terms to get the relevant documents to the query. Documents are then ranked using 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 (2012a, b) which proves that the cosine measure is the most efficient one in comparison to other statistical measurements in order to rank the documents according to the most relevant to the user's query. The cosine measure calculates the angel between two documents (between a document and the user's

query which is treated as a document) representation vectors. Thus, a cosine value of zero means that the query and document vectors were orthogonal to each other and means that there is no match or the term simply did not exist in the document being considered. To know the cosine relations between two documents (documents D and querry Q) (show in Eq. 1 Q):

$$\text{Cosine}(D,Q) = \frac{|D \cap Q|}{\sqrt{|D| \times |Q|}}$$

Where:

Cosine (D, Q) = The cosine similarity relationship between document D and a user's query Q
D = The document in the collection
Q = The user's query

After calculating the similarity measure, the ranked list appears to the user as the answer to his/her query. This list is arranged from the highest values of the cosine measure to the lowest one as a weight as a ranked list.

Module 5 (Usage-based re-ranking): This module is a re-ranking process based on several parameters like: frequency of pages visited by the user, the time spent on those pages and click history of the pages, the information of these parameters is taken from log files after analysis as presented in log files analysis's section; also these log files are kept by web servers. The parameters include time spent on the page, frequency of visits and for more explanation see usage-based parameter's section; In addition, considering also the ranking list of pages that comes from applying the content-based algorithm. This new algorithm is a Hybrid Usage-Based Ranking Algorithm (EHURA), EHURA's idea is based on the combination of usage-based parameters and the pervious ranking list, their combination provided a new weight for pages in order to re-ranking them as a new ranked list that appears to the user.

Hybrid Affix Removal algorithm: Arafat and Saad (2008) presented a hybrid Arabic stemmer 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. The algorithm of the Hybrid Affix Removal algorithm is as the following.

Suffix removal:

- For words ending with a suffix and of at least 3 characters without the suffix
- Replace the suffix with "Ha" only if it is "Att" and the word isn't starting with "Al" or its length>5 letters, otherwise remove "Att". If so, stop searching for more suffixes
- Replace the suffix by "ya" if it is "Ahh"
- Stop searching for Suffixes if the word starts with "Al" and is having any suffix except "Ha" and "ya" or its length without the suffix is ≤5 characters
- Otherwise, remove the suffix and stop searching for suffixes

Prefix removal

Phase 1:

- For words starting with "Alf Alf remove Alf"
- For words of at least 5 characters starting with "wal", remove "wal " and stop searching
- For words of >4 characters, starting with "lell ",remove "lell " and stop searching
- For words of at least 4 characters, starting with "Al" then remove "Al", mark it as a non stop word and stop searching for prefixes

Phase 2:

- For words of at least 3 characters, starting with a prefix, remove the prefix re-check for stop words and remove them

Log files analysis: Web servers collect large volumes of web usage data. This data is stored in web log files. This usage data is important in the usage-based parameters which are taken from log files after analysis. This study explains the log file analysis. Really, the log file contains lot 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), file does not exist, server internal error, service temporarily unavailable, etc., before retrieving the usage data. Log file analysis consist a series of processes such as data cleaning, user identification, session identification as in the following.

Data cleaning is the process of removing unnecessary records like graphics, video and formatted information like css. In addition, this process removes the records of failed HTTP status codes. User Identification is the process of identifying users and user agent fields of log entries, its considered as the following:

- Different IP addresses refer to different users
- The same IP with different operating systems or different browsers should be considered as different users
- While the IP operating system and browsers are all the same, new users can be determined as to whether the requesting page can be reached by accessed pages before according to the topology of the site
- A user session is considered to be all of the page accesses that occur during a single visit to a web site. Session identification is the process of defining users that may access the site more than once

Usage-based parameters: In this stage, the system calculates two usage-based parameters as in the following:

Frequency of visit: Determines the relevance of a web page by its selection frequency in order to find the frequency weight which is the admittance frequency of a page u is the number of times the page is visited and the page rank which appears in the ranked list from the previous stage. The frequency weight in Eq. 2 is:

$$FW = \frac{\text{Number of visit on a page (u)}}{\text{Total number of visit on all page}} \times PR(u) \quad (2)$$

Where:

FW = Frequency weight
PR (u) = The page rank of a page u

Time spent: It shows how long the users spend on a page after removing the download time of the page. Because a user generally spends more time on the useful pages and does not waste more time on screening the page and rapidly skipping to another page. So, it's an important parameter to indicate the usefulness of the pages, this parameter is considered to calculate the real time spent on a page by taking the value of time spent on the page from the log file, subtracting from the download time in order to find the time spent weight. It is calculated as follows:

$$TW = \frac{\text{Time spent on a page (u)-Download time (u)}}{\max(\text{Time spent on a page (u)-Download time (u)})} \quad (3)$$

where, TW, time spent weight:

$$\text{Dowload time (u)} = \frac{\text{Size of a page (u)}}{\text{Transfer rate for page (u)}} \quad (4)$$

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

$$CTW = \{0.5 \text{ if an event is done} | 0 \text{ otherwise}\} \quad (5)$$

where, CTW, Click-Through Weight. The combination (i.e., the summation) of the above parameters (frequency of visit, time spent and click-through) with the content-based ranking results, provide our EHURA algorithm which is trying to improve the ranking result from Arabic web search engines.

Performance analysis: In order to study the performance of the proposed system we used different evaluation measures. These measures are discussed in next section. Then, the data sets used and the experimental results are shown in experimental result's section.

Evaluation of the proposed 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 result with the Content-Based Ranking algorithm. The performance measured by the recall and precision measurements and other measures are represented in the following formulas:

$$\text{Precision} = \frac{|\{\text{relevent documents}\} \cap \{\text{retrieved documents}\}|}{\{\text{retrieved documents}\}} \quad (6)$$

$$\text{Recall} = \frac{|\{\text{relevent documents}\} \cap \{\text{retrieved documents}\}|}{\{\text{relevent documents}\}} \quad (7)$$

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

$$\text{Fall-out} = \frac{|\{\text{non-relevent documents}\} \cap \{\text{retrieved documents}\}|}{\{\text{non-relevent documents}\}} \quad (8)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

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

$$\text{Avep} = \frac{\sum_{i=1}^{N_0} P_i(r)}{N_q} \quad (10)$$

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

RESULTS AND DISCUSSION

For testing the proposed system, it was applied on Ain Shams Arabic corpus. This corpus belongs to the modern standard Arabic type; it contains 242 documents with different sizes and tested the system 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 and it was compared with other information retrieval system called Content-Based Information retrieval System (CBIS) in order to prove its effectiveness; the CBIS is based on the Content-Based Ranking Algorithm (CBRA). Figure 2 shows the precision and recall results for each query for the Content-Based Ranking Algorithm (CBRA) of CIBS in comparison with the Hybrid algorithm (EHURA) in the

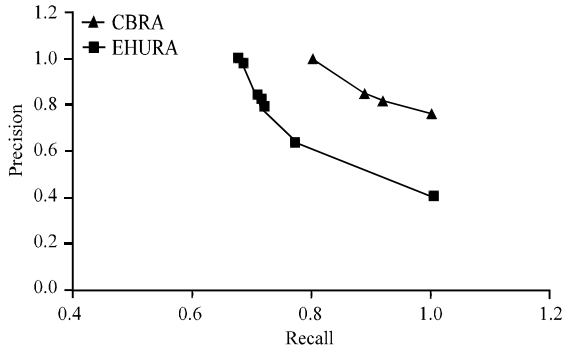

Fig. 2: Precision and recall for ranking against the Arabic Ain Sham's corpus 20 queries

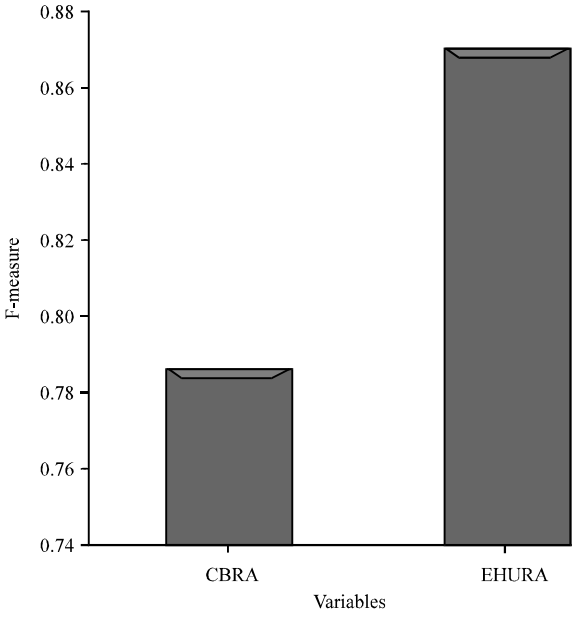

Fig. 3: F-measure for ranking against the Arabic Ain Sham's corpus

Table 1: Evaluation summary for the proposed system

Variables	Precision	Recall	Fall-out	F-measure
CBRA	0.8753	0.7125	0.2875	0.7856
EHURA	0.9802	0.7800	0.2200	0.8687

proposed system. It's clear that EHURA reach a better result than the content-based one. The average precision of our new approach (EHURA) reached 98% while the precision of the Content-Based Ranking Algorithm (CRBA) is 88%, the results are shown in Table 1. So, our proposed EHURA algorithm improves the precision over the Content-Based Ranking Algorithm (CRBA) 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 (CRBA) 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 Fig. 3 that the EHURA algorithm improved the F-measure over the Content-Based Algorithm (CRBA) by 8%.

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 were done to improve the ranking algorithm for web search engines however, there are still several drawbacks. Rarely of these researches are done upon Arabic language search engines. Thus in this study we proposed an efficient Arabic information retrieval system using a new Hybrid Usage-Based Ranking algorithm called EHURA. The objective of this algorithm is to overcome the drawbacks of the ranking algorithms and improve the efficiency of web searching.

The system was applied to Ain Shams Arabic corpus for testing and evaluation. The results show that the EHURA 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 (CBRA) by about 10% while improving the recall percentage by 7%.

REFERENCES

Anonymous, 2015. Internet world stats: Usage and population statistics. Miniwatts Marketing Group, USA. http://www.internetworldstats.com/stats.htm.

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.

Ding, C., C.H. Chi and T. Luo, 2002. An improved usage-based ranking. Proceedings of the International Conference on Web-Age Information Management, August 11-13, 2002, Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-44045-1, pp: 346-353.

Hajeer, S., 2012a. Vector space model: Comparison between euclidean distance and cosine measure on Arabic documents. Int. J. Eng. Res. Appl., 2: 2085-2090.

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

Hajeer, S.I., R.M. Ismail, N.L. Badr and M.F. Tolba, 2015. An efficient hybrid usage-based ranking model for information retrieval systems & web search Engine. Proceedings of the 6th International Conference on Information and Communication Systems (ICICS), April 7-9, 2015, IEEE, Amman, Jordan, ISBN:978-1-4799-7349-1, pp: 142-147.

Hajjar, M., K. Zreik and A. Al-Hajjar, 2010. The search engine performance in the Arabic language. Proceedings of the International Conference on Society for Information Technology & Teacher Education, March 29, 2010, AACE, Chesapeake, Virginia, ISBN:978-1-880094-78-5, pp: 2239-2245.

Jain, R. and D.G. Purohit, 2011. Page ranking algorithms for web mining. Int. J. Comput. Appl., 13: 22-25.

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.

Kritikopoulos, A., M. Sideri and I. Varlamis, 2007. Success index: Measuring the efficiency of search engines using implicit user feedback. Proceedings of the 11th Conference on Pan-Hellenic on Informatics, Special Session on Web Search and Mining, May 18-20, 2007, Athens University of Economics and Business, Greece, pp: 1-14.

Lebbos, G., K. Zreik and M. El-Sayed, 2014. Performances of the most popular search engines in Arabic language. Intl. J. Comput. Theory Eng.,6: 4-8.

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.

Meftouh, K., M.T. Laskri and K. Smaili, 2010. Modeling arabic language using statistical methods. Arabian J. Sci. Eng., 35: 69-82.

Mukherjee, I., V. Bhattacharya, S. Banerjee, P.K. Gupta and P.K. Mahanti, 2012. Efficient web information retrieval based on usage mining. Proceedings of the 1st International Conference on Recent Advances in Information Technology (RAIT), March 15-17, 2012, IEEE, Dhanbad, India, ISBN: 978-1-4577-0694-3, pp: 591-595.

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

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