American Journal of Applied Sciences 10 (9): 1018-1024, 2013
ISSN: 1546-9239

doi:10.3844/ajassp.2013.1018.1024 Published Online 10 (9) 2013 (http://www.thescipub.com/ajas.toc)

Evaluation of Different Query Expansion Techniques for Arabic Text Retrieval System

[1]Aysh Alhroob, [2]Hayel Khafajeh and [3]Nisreen Innab

[1]Department of Software Engineering, Al-Isra University, Amman, Jordan
[2]Department of Computing, Zarqa University, Zarqa, Jordan
[3]Department of Science and Technology, University of New England, Australia

Received 2013-01-08, Revised 2013-03-13; Accepted 2013-07-31

ABSTRACT

The word mismatch problem is fundamental to Information retrieval. Query expansion process helps to overcome this problem. Based on the Arabic corpuses, the comparisons between two query expansion techniques (global and local query) have been conducted to determine the query effectiveness. First one represents the local context analysis which represents a local method, while a global method was the second technique that has been represented by the Association and similarity thesauruses. These techniques can be used in any special field or domain to improve the expansion process and to get more relevant documents for the user's query. This study introduces a comparison between these approaches and shows their effectiveness. Although, local context analysis has some advantages over the similarity thesaurus, Association thesaurus which is global is generally the most effective one.

Keywords: Query Expansion, Association Thesaurus, Similarity Thesaurus, Local Context Analysis, Thesaurus, Indexing, Natural Language (NL), Synonyms

1. INTRODUCRTION

Millions of users search daily through the internet and other information stores. They search for their needs of information by writing their queries. Unfortunately, these queries may fail to reach to their needs. This failure refers to the different words that used by the searchers in their queries and the words that authors used to describe the same concepts of their documents. This is known as word mismatch. Furthermore, other minor problem is the numbers of the query words normally are very short. As an example, applications that provide searching across the World Wide Web typically record average query lengths of two words (Tawileh *et al*., 2010). Query expiation is one of the obvious approaches that used to solve these problems, since these problems tend to decrease when queries gets longer. Both local and global techniques based on the query words have the advantages of expanding that query, because expanding the query by adding new words to the query with similar meaning increases the matching words chance and will get more relevant documents. The main idea behind using thesaurus in the information retrieval is query expiation; however, a general thesaurus of any use has a little evidence in improving the effectiveness of the search (Tamine-Lechani *et al*., 2010). On the other hand, Local Context analysis expanded the query by adding the expansion words from the relevant documents. Firstly, the query is written and the information retrieval process is conducted. Then, the terms are chosen to expand the query only from the top ranked documents which assumed to be relevant are considered to expand the query. Use of context and phrase structure is an idea that borrowed from global analysis technique, but we applied them to the local document set. In this study, 242 Arabic abstract documents have been selected. These documents were presented at the Saudi Arabian National Computer Conference in

Corresponding Author: Aysh Alhroob, Department of Software Engineering, Al-Isra University, Amman, Jordan

addition to 59 Arabic queries. All these abstracts involve computer science and information system. In this study, an automatic information retrieval system has been introduced from scratch to handle Arabic data.

2. BACKGROUND AND PREVIOUS WORK

Several approaches to query expansion have been studied in the past. However, more recently, attention has been focused on techniques that analyse the entire document corpus to discover word relationships (global analysis) and those that limit the analysis to documents retrieved by the initial query (local analysis). The term mismatch problem between user queries and documents has been introduced by several previous works. One of the earliest studies was carried by (Jones, 1971) who used clustered words based on co-occurrence in documents to expand the query. Then, global analysis and local analysis techniques are used by (Imran and Sharan, 2009).

2.1. Global Query Expansion

Through global query expansion (Lixin and Guihai, 2009) the query is expanded depending on extract information from all the documents in the database, if these documents are related to the query or not. Some of global query expansion techniques.

2.2. Thesaurus

The thesauri are built manually or automatically (Liang-Yu and Shyi-Ming, 2007) Building thesaurus manually requires the study of words meanings and the relation between these words according to the meaning like synonyms and antonyms. These studies are expensive, need a lot of time and effort and suffer from the bias problems (Liang-Yu and Shyi-Ming, 2007). Although the manual thesauri are characterized by the difficulty of building them, are of good quality because the relations between words are built by specialists. The automatic methods to build the thesaurus are characterized by the high precision in the determination of relations between words and the possibility of using the same method for more than one language and a great number of corpuses can be used to build the automatic thesaurus (Nakayama *et al*., 2007). However, the automatic thesauri encounter a problem which is the difficulty of the results' evaluation. The thesauri are widely used in query expansion through adding new words to the query before beginning with the retrieval process. The thesauri are different according to the kind of relation between the words that the thesaurus depends on such as the statistical relations between words as in the similarity thesaurus or the relation between words according to meaning (Reinhard, 2012). The thesaurus represents an important source for many researches that are dealing with natural language processing such as the researches specialized in information retrieval. The thesauri are used to overcome many problems, which encounter the access to information and its retrieval. The thesaurus is a collection of terms plus the assets of relations among them (Banko *et al*., 2009). Each thesaurus connects each word with a group of words with a relation or a number of relations which are determined when the thesaurus is designed.

The Phrase Finder is suggested by (Liang-Yu and Shyi-Ming, 2007) in this research the association thesaurus used through which the term is connected with the phrase and ignored the other terms. Despite the fact is the phrase has one term or more. The thesaurus is built through co-occurrence between phrases and terms and the related phrases are used to expand the query to improve the retrieval process. After the writing of the user's query, the outcome is an ordered list of phrases. The query is expanded by adding the high ranked phrases in the previous list.

Hidetsugu (2007) introduced a new method to build the thesaurus in English language and Japanese language. They also define four kinds of relations between language words to build these thesauri. The first of these relations is: Hypernym/Hyponym: Used to extract the terms that are related by the Hypernym/Hyponym relation. The second relation is a defining what they call abbreviation extraction. It depends on the relation between the terms and their abbreviations. The third relation is what they called synonym extraction is. To find these kinds of relations, the researchers concentrate on the citation relation between terms. The fourth kind of relations is what they call related terms extraction. In the results they presented that the suggested system has improved the information retrieval process through query expansion.

Senellart and Blondel (2002) a method to extract synonyms of the dictionary has been proposed. The method depends on the similar words that use a number of definitions and these words that are used in the definition are also used to define other words. This means, if the words C and D are used to define words A and B, thus A and B are similar.

Egozi *et al*. (2011) presented a query expansion method that depends on the mixture between the global method and the local method by building a local thesaurus from the resulted documents of the user initial query. The user writes his query, then the documents with high rank and that are related to the user's query are taken to build a thesaurus by using these documents and then the query is expanded depending on the previous thesaurus.

2.3. Local Query Expansion

In the local query expansion, the query is expanded by adding the expansion words from the relevant documents. Firstly, the query is written and the information retrieval process is conducted. Then, the high-rank documents are taken to expand the query through them. In a local strategy, the top-ranked documents retrieved for a given query are examined to determine terms of query expansion.

The relevance feedback process introduced in the mid of 1960s (Carpineto and Romano, 2012). The relevance feedback is used to improve the user's query. The initial search is conducted through the system and using the users' query. The system retrieves a set of the ranked documents and then the user determines which of these documents is relevant to his query. The system reforms the query based on the user judgment (the determination of the relevant and irrelevant documents). The system repeats the retrieval process by using the modified query. This process is still repeated until the user gets the suitable documents for him.

In the previous work, Relevance judgments are used to estimate the probability of a term related to another term or query (term classification) or to estimate the document's relationship to other documents (document clustering). These approaches are impractical because relevant judgments are not often available and, even if available; relevance judgments are often produced for a set of index terms or a particular query, which do not cover a whole collection. In addition to use the term selection for query expansion is based on individual terms (local analysis) instead on the collection of whole index terms (global analysis).

Local context concepts are selected based on co-occurrence with query terms, concepts are chosen from the top ranked documents (Tamine-Lechani *et al*., 2010) and the best passages are used instead of whole documents. Local context analysis involves only by the top ranked documents that have been retrieved by the query, i.e., the top ranked documents for a query were proposed as a source of information, so the most frequent 20 terms and 10 phrases (none stop words) from the top ranked are added to the query.

3. COMPARISON

This study shows the comparison among three query expansion techniques, two of them-similarity thesaurus and association thesaurus-represent the global query expansion technique. The third one is local context analysis represents the local query expansion techniques.

Nidal *et al*. (2010) compared between local context analysis and similarity thesaurus using Arabic corpus. According to the researchers, the results of both techniques enhance the retrieval process. The results also showed that the similarity thesaurus outperforms the local context analysis technique. In the study (Khafajeh *et al*., 2010), the researchers used Arabic corpus with two global query expansion techniques: The similarity thesaurus and the association thesaurus. They found that both of the used techniques provide good results compared with traditional information retrieval systems. The researchers' experiments showed that the association thesaurus outperforms the similarity thesaurus.

An association thesaurus has been constructed by defining a term-term correlation matrix whose rows and columns are associated to the index term in the document collection. In this matrix, a normalized correlation factor $c_{i,l}$ between two terms ki and kl has been defined by using the formula:

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_i - n_{i,l}}$$

Where:

ni = The number of documents, which contain the term (ki) (Index Terms)
n_l = The number of documents, which contains the query terms (k_l)
n_{il} = The number of documents, which contains both terms (k_i) and (k_l))

Authors used the term correlation matrix to define a fuzzy set associated to each index term k_i. In this fuzzy set, a document dj has a degree of membership as it proposed by Alzahrani *et al*. (2012) and Kobayashi in the fuzzy set model:

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$$

Where:

$\mu_{i,j}$ = The membership of document d_j in the fuzzy set association with term k_i

K_l = Index term in document d_j

C_{il} = Correlation factor between (term k_i) and k_l terms in document d_j

In the end we present the results of the comparison technique that show the advantage of similarity thesaurus and association thesaurus and local query expansion.

4. EXPERIMENTS

The experiments are demonstrated of global, local query expansion techniques using full words and stemmed words. The query was expanded by adding two words for each word of the query's words by using the two global techniques, the Similarity thesaurus retrieving systems using the Vector Space Model (VSM) VSM, which depends on Cosine measure and the Association thesaurus using a fuzzy set model. While the experiments of the Local Context Analysis were implemented to expand the query by adding two words for each word, through using the first ten passages and ten words to determine the passage.

These results shows that contained the retrieved documents in ascending order according to their similarity or the relationship values. The three information retrieval systems can be browsed in the following way.

4.1. Experiment 1

Table 1 and Fig. 1 show clearly the improvement of using an Association thesaurus over Local Context Analysis (LCA) and using the LCA over using Similarity thesaurus, all of that were in the case of using full words retrieving.

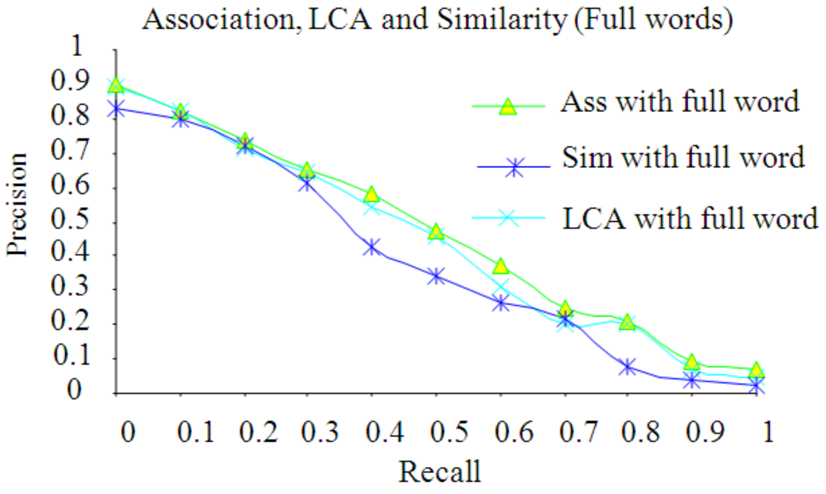

Fig. 1. A comparison between the values of average Recall Precision when full words were used

Table 1. Averages of retrieving information in the case of using Full words through using LCA technique, Similarity and Association thesauri

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Association with full word	0.90	0.82	0.74	0.65	0.58	0.47	0.37	0.25	0.21	0.09	0.07
LCA with full word	0.89	0.82	0.71	0.64	0.54	0.46	0.31	0.20	0.20	0.07	0.05
Similarity	0.83	0.80	0.72	0.61	0.43	0.34	0.26	0.22	0.08	0.04	0.02

Table 2. Averages of retrieving information in the case of using stemmed words

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Association with full stemming	0.955	0.89	0.86	0.77	0.69	0.59	0.51	0.40	0.35	0.24	0.13
LCA with stemming	0.925	0.88	0.79	0.75	0.64	0.53	0.47	0.31	0.23	0.12	0.09
Similarity with stemming	0.901	0.83	0.73	0.68	0.59	0.47	0.42	0.26	0.20	0.10	0.06

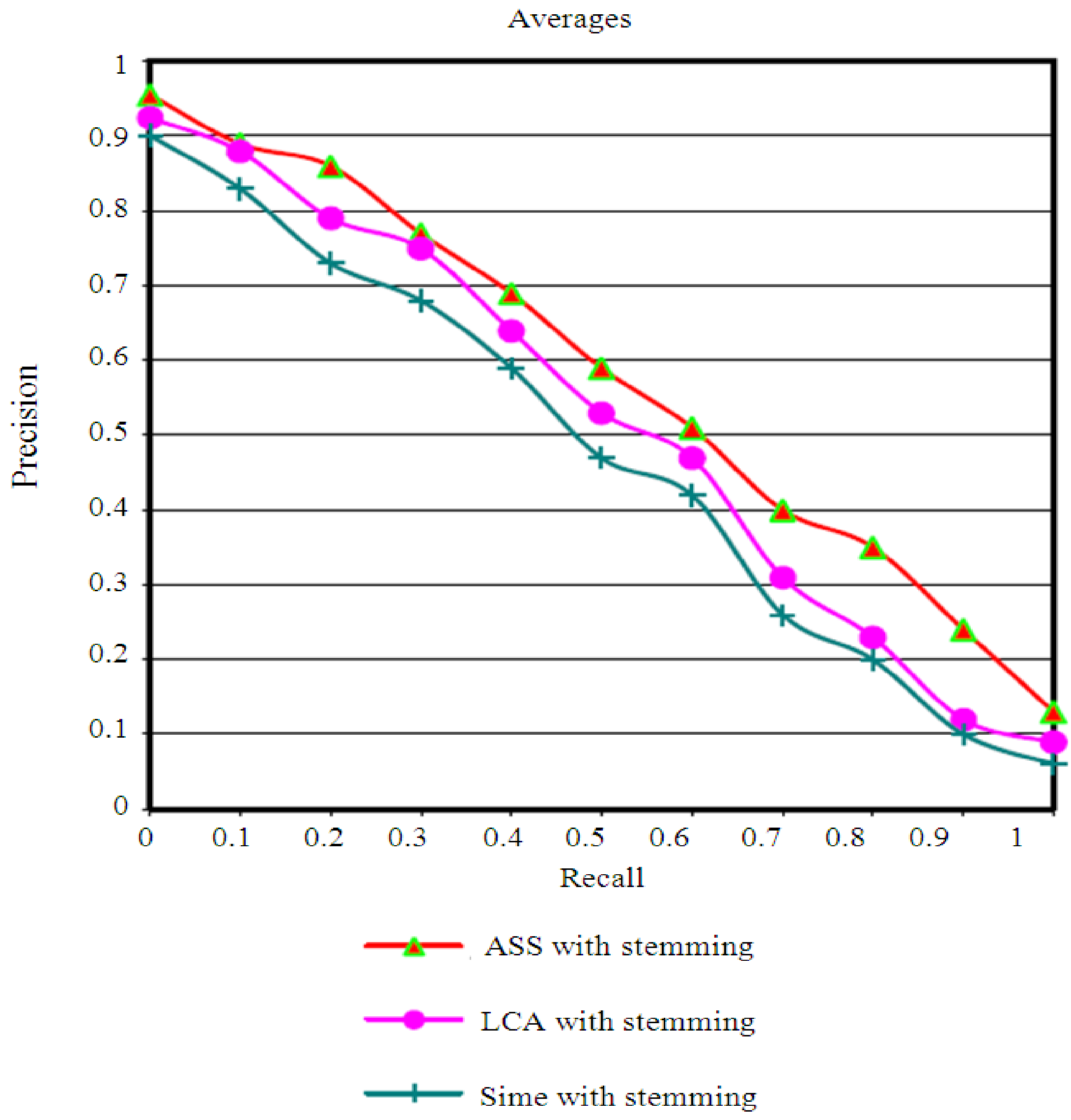

Fig. 2. A comparison between the values of average Recall Precision when stemmed words were use

The chart in **Fig. 1** shows the effect of using thesauri on the criterion of average recall precision. When Association thesaurus was used, the results were better than using LCA, while using the Association thesaurus and LCA technique were much better than using similarity thesaurus.

4.2. Experiment 2

Table 2 and **Fig. 2** show the results of using stem words retrieving, where they clearly show that the Association thesaurus got the best result over the other two cases.

The chart in **Fig. 2** shows the effect of using an association thesaurus on the system efficiency that depends on the stemmed words better by applying the criterion of average recall precision. When association thesaurus was used, the results were the best, while using the LCA technique with stemmed words was better than using Similarity thesaurus, which agreed with the case of using full words.

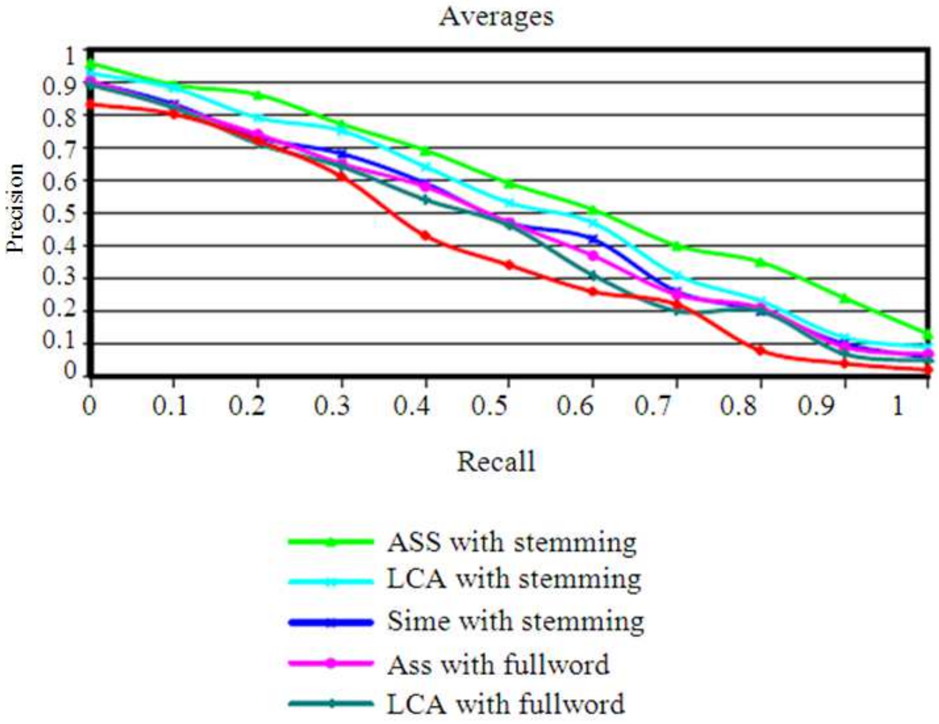

Fig. 3. A comparison between the values of average Recall Precision when stemmed words and full words were use

Table 3. Average of all the Relative work

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Ass with stemming	0.955	0.89	0.86	0.77	0.69	0.59	0.51	0.40	0.35	0.24	0.13
LCA with stemming	0.925	0.88	0.79	0.75	0.64	0.53	0.47	0.31	0.23	0.12	0.09
Simi with stemming	0.901	0.83	0.73	0.68	0.59	0.47	0.42	0.26	0.20	0.10	0.06
Ass with full word	0.90	0.82	0.74	0.65	0.58	0.47	0.37	0.25	0.21	0.09	0.07
LCA with full word	0.89	0.82	0.71	0.64	0.54	0.46	0.31	0.20	0.20	0.07	0.05
Similarity	0.83	0.80	0.72	0.61	0.43	0.34	0.26	0.22	0.08	0.04	0.02

4.3. Experiment 3

Table 3 shows the effect of using the stemmed words for information retrieving was always better than using Full words, This goes well with (Khafajeh *et al*., 2010) when he said that using the roots of the words in Arabic will make the efficiency of the Arabic IRS better and it shows that using other techniques such as LCA got an improvement over using similarity thesaurus, which leads to say that using thesauri as in (Khafajeh *et al*., 2010) or LCA as in (Nidal *et al*., 2010) is much better than using similarity and traditional information retrieval. The best case was about using stemming words and the use the Association thesauruses with the stemmed words was the best case over all the other cases.

In the next graph in **Fig. 3** we can conclude that using of Association thesaurus with the stemmed words is the best case over all the other cases.

5. CONCLUSION

This study shows that the stemmed retrieval methods performed significantly better than the full word retrieval method, which agreed with all other studies in the different languages. At the same time, using the Association thesaurus in the Arabic language retrieving system is much better than using Similarity thesaurus, which agreed with (Khafajeh *et al*., 2010). Furthermore, provides a better retrieval performance than using Local Context analysis. Based on our result, using Local

Context analysis in the Arabic language retrieving system is much better than using Similarity thesaurus, which agreed with (Nidal *et al*., 2010). Whereas, there is a possibility for applying automatic indexing and it's equations in the Arabic language. It is good to use the stemming of Arabic words reinforces and supports IRS for other languages, as it agreed with (Kanaan and Wedyan, 2006; Tawileh *et al*., 2010). The best results were gained when both the stemming and the Association thesaurus were used together. While the worst results were found when no stemming words and the similarity thesaurus were used together.

This study could be applied to other different documents and techniques such as co-occurrence of the terms. The user can be utilized in feeding back the system in order to have a high precision thesaurus; here the user can interfere in choosing the words to widen the query. This is to increase the degree of similarity between this new word and the original one and decreasing the similarity between the word he chooses and those unneeded. On the other hand, this study may lead researchers to improve and enhance an algorithm to build query automatically.

6. REFERENCES

Alzahrani, S.M., N. Salim and A. Abraham, 2012. Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Trans. Syst. Man Cybernet., Part C. Appli. Rev., 42: 133-149. DOI: 10.1109/TSMCC.2011.2134847

Banko, M., M.J. Cafarella, S. Soderland, M. Broadhead and O. Etzioni, 2009. Open information extraction from the web. University of Washington.

Carpineto, C. and G. Romano, 2012. A survey of automatic query expansion in information retrieval. ACM Comput. Surveys, 44: 1-56. DOI: 10.1145/2071389.2071390

Egozi, O., S. Markovitch and E. Gabrilovich, 2011. Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inform. Syst., 29: 8-50. DOI: 10.1145/1961209.1961211

Hidetsugu, A., 2007. Query expansion using an automatically constructed thesaurus. Proceedings of the NTCIR-6 Workshop Meeting, May 15-18, Tokyo, Japan, pp: 414-419.

Imran, H. and A. Sharan, 2009. Thesaurus and query expansion. Int. J. Comput. Sci. Inform. Technol., 1: 89-97.

Jones, K.S., 1971. Automatic Keyword Classification for Information Retrieval. 1st Edn., Butterworths, London, ISBN-10: 0408701374, pp: 253.

Kanaan, G. and M. Wedyan, 2006. Constructing an automatic thesaurus to enhance Arabic information retrieval system. Proceedings of the 2nd Jordanian International Conference on Computer Science and Engineering, (SE' 06), Salt, Jordan, pp: 89-97.

Khafajeh, H., N. Yousef and G. Kanaan, 2010. Automatic query expansion for Arabic text retrieval based on associated and similarity thesaurus. Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems, (IS' 10), pp: 1-17.

Liang-Yu, C. and C. Shyi-Ming, 2007. A new approach for automatic thesaurus construction and query expansion for document retrieval. Inform. Manage. Sci., 18: 299-315.

Lixin, H. and G. Chen, 2009. HQE: A hybrid method for query expansion. Expert Syst. Appl., 36: 7985-7991. DOI: 10.1016/j.eswa.2008.10.060

Nakayama, K., Hara, T. and S. Nishio, 2007. A thesaurus construction method from large scaleweb dictionaries. Proceedings of the 21st International Conference on IEEE Advanced Information Networking Applications, May 21-23, IEEE Xplore Press, Niagara Falls, ON, pp: 932-939. DOI: 10.1109/AINA.2007.23

Nidal, Y., I. Al-Bidewi and M. Fayoumi, 2010. Evaluation of different query expansion techniques and using different similarity measures in Arabic documents. Eur. J. Scientific Res., 43: 156-166.

Reinhard, R., 2012. The automatic generation of thesauri of related words for English, French, German and Russian. Int. J. Speech Technol., 11: 147-156. DOI: 10.1007/s10772-009-9043-7

Senellart, P.P. and V.D. Blondel, 2002. Automatic discovery of similar words. Survey Text Mining, 2: 25-44.

Tamine-Lechani, L., M. Boughanem and M. Daoud, 2010. Evaluation of contextual information retrieval effectiveness: Overview of issues and research. Knowl. Inform. Syst., 24: 1-34. DOI: 10.1007/s10115-009-0231-1

Tawileh, W., J. Griesbaum and T. Mandl, 2010. Evaluation of five web search engines in Arabic language. University of Hildesheim.