Query Expansion based-on Similarity of Terms for Improving Arabic Information Retrieval

Khaled Shaalan[1], Sinan Al-Sheikh[2], Farhad Oroumchian[3]

[1]The British Univ. in Dubai, PO BOX 345015, Dubai, UAE
Khaled.shaalan@buid.ac.ae
[2]IBM, Dubai Internet City, PO BOX 27242, Dubai, UAE
sinan@ae.ibm.com
[3]University of Wollongong in Dubai, POBox 20183, Dubai, UAE
FarhadOroumchian@uowdubai.ac.ae

Abstract. This research suggests a method for query expansion on Arabic Information Retrieval using Expectation Maximization (EM). We employ the EM algorithm in the process of selecting relevant terms for expanding the query and weeding out the non-related terms. We tested our algorithm on INFILE test collection of CLLEF2009, and the experiments show that query expansion that considers similarity of terms both improves precision and retrieves more relevant documents. The main finding of this research is that we can increase the recall while keeping the precision at the same level by this method.

Keywords: Arabic, Arabic NLP, Arabic Information Retrieval, Query Expansion, EM algorithm

1 Introduction

Information Retrieval (IR) is the process of finding all relevant documents responding to a query from unstructured textual data. The traditional model for IR assumes that each document is represented by a set of keywords, so-called index terms. An index term is simply a word whose semantics contributes to the document topic. The challenge increases when the number of documents stored grows, the content carries different topics, few words are used in queries, and more clarifications about words in queries are needed.

Arabic language has a very rich set of vocabulary, which with their synonyms introduce a problem to the IR process [15-16]. Many synonyms can contribute to the same meaning of the sentence. An example that shows the challenge in IR using synonyms is the query for "كأس العالم" (the World Cup) which could miss documents represented by the keyword "مونديال", (borrowed from the French "Mondial"). With this

set of vocabulary, it is not very hard to write an entire essay in Arabic about, say, how sports benefit human health, and yet do so without ever using the keywords 'Sports', 'Human', or 'Health'.

On top of the previous challenge users tend to input very limited set of words as their intended query. Many researchers such as Stefan Klink [8] have indicated that the average words used in a query is around two to three words. It is a challenge to hit users' real need for information using very limited number of words especially when those few words might carry different meanings, like the case in Arabic. Both the limited set of words in users' query and the potential absence of words from this set in the retrieved documents are the motivation behind this research. Query expansion is a proposed solution to overcome those two problems and successfully retrieved documents that were previously over looked.

Query expansion is considered as a Meta-level process that is used to add more information to clarify the user's query. It is the process of rebuilding new informed queries from an existing one in order to improve the retrieval performance and help in matching additional documents. Many query expansion techniques can be used. They are classified into two categories: automatic expansion based on linguistic knowledge and semi-automatic based on user feedback. In [1], it demonstrates an approach where a query is expanded by adding more synonyms. Whereas, in [2] it shows how a query is expanded by stemming its terms and adding common suffixes and prefixes. Semi-automatic expansion algorithms have been used to add/rebuild input query from user feedback like Probabilistic Relevance Feedback (PRF) expansion algorithm [4-5]. Essentially that algorithm Compares the frequency of occurrences of a term in documents that user marked as relevant with terms in the whole document collection. So if a term occurs in the documents marked as relevant more frequently than in the whole document collection it will be assigned a high weight.

The proposed technique depends on the co-occurrence of words while expanding queries. A paragraph about "كأس العالم" (world cup) uses common words such as " كرة القدم"(football), "كرة"(ball), "أهداف"(goals), "حماسة"(excitements), "كأس البطولة"(championship cup) ... etc. Those words are also present in the documents that do not have the exact match of keywords "كأس العالم" (world cup) however, it has the word "مونديال" which also means (world cup).

The proposed technique starts by analyzing documents that have the exact wording of the query in order to identify a list of co-occurring contextual words. This list of words will be used to expand the current query. The expanded query will then be used to pull other documents. New set of documents do not necessarily have the exact words as the original query. This way, it was possible to expand a query based-on similarity of terms for improving Arabic Information Retrieval.

The remaining of this paper is structed as follows. Section 2 presents a background appraisal showing other people work in this area. Section 3 gives a detailed description of the proposed algorithm. Section 4 explains how EM was used to optimize and improve query expansion. Section 5 shows the testing experiments we conducted and points of improvement that the proposed solution provided. Finally section 6 concludes and sums up the main findings.

2 Background

Matthew W. Bilotti [11] discussed "Query Expansion Techniques for Question Answering" in his thesis. He discussed five query expansion techniques, two term expansion methods and three term-dropping strategies. His results show that there are well-performing query expansion algorithms that can be experimentally optimized for specific tasks.

Hayel Khafajeh, and others [12] compare the performance of search engine before and after expanding queries. Their approach to expand queries was based on Interactive Word Sense Disambiguation (WSD). They found that expanding polysemous query terms by adding more specific synonyms will narrow the search into the specific targeted request and thus causes both precision and recall to increase; on the other hand, expanding the query with a more general (polysemous) synonym will broaden the search which would cause the precision to decrease. Their method of expanding queries depends on user feedback for the results.

Hayel Khafajeh and others [9] also worked on automatic Arabic thesauri that can be used in any special field or domain to improve the expansion process. Their efforts concluded that the association thesaurus improved the recall and precision over the similarity thesaurus. However, it has many limitations over the traditional information retrieval system in terms of recall and precision level.

T. Rachidi, and others [10] depended extensively on Arabic root extraction to build expanded queries. They also relied on three concept thesauri while expanding their queries. The first one was built manually, the second was built automatically from crawled XMLs documents and the last one was built automatically from an automatic categorization for the crawled documents. They reported a %75 improvement on their first experiment using query expansion.

3 The Proposed System

Fig. 1 shows the structure of our Arabic information retrieval system. It consists of two fundamental stages: indexing and querying. The indexing stage handles pre-processing, term selection and indexing. In the indexing stage, we also build a relationship database which is useful in runtime processing of top 10 documents and extracting their best terms. The querying stage handles pre-processing of user queries and retrieval of relevant documents using the previously indexed text. The two stages are interdependent. Indexing is designed such that it facilitates querying.

The contribution of this paper lies in the way we are expanding the original query. The indexing stage consists of three steps. It starts by passing Arabic documents into a pre-processing and noise removal phase. The data pre-processing and noise removal step takes care of cleaning up the Arabic text. The tasks performed by this step include: Duplicate white spaces removal, excessive tatweel (or Arabic letter Kashida) removal, HTML tags removal, and Handling special characters (i.e. {!@#$^%&*;+:()_}).

Following the data pre-processing and noise removal step, Arabic documents are passed into baseline indexing and relationship database construction in parallel. For the baseline indexing, the system parses Arabic documents and stems its words before it produces the baseline index. The relationship database construction builds the relationship database which is an SQL database that links each word in a document to all documents that it has occurred in.

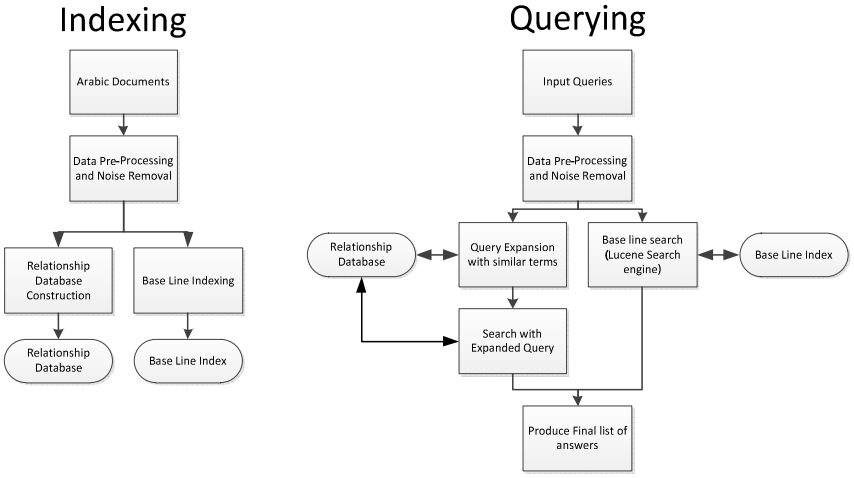

Fig. 1. Arabic Information Retrieval System Architecture

The querying stage consists of five steps. It takes both the baseline index and the relationship database as input. This stage handles the pre- and post-processes performed on queries fed to the system. The first step of the system is pre-processing of the input queries. The pre-processing follows the same steps for both queries and documents. In addition to that it also cleans the query from all stop words. Query processing proceeds with parallel steps: baseline search and query expansion with similar terms. The baseline step applies the baseline retrieval approach and retrieves the baseline documents. The query expansion component expands the query with similar terms and then retrieves another set of documents using expanded query. The details of the query expansion step are described in the next section.

Finally, the last step is to combine both retrieved sets of documents. While merging two lists of documents duplicate answers will be promoted.

4 Query Expansion

Query expansion is the process of adding extra data to the input query in order to provide more clarity. Expanding queries in this work consists of three steps:

- Extracting top 10 documents
- Extracting top 100 keywords out of the top 10 documents, and
- Eliminating irrelevant keywords using EM distance.

The remaining words are then added to the original query to construct the expanded version of the query. Fig. 2 represents the steps followed when expanding a query in this paper:

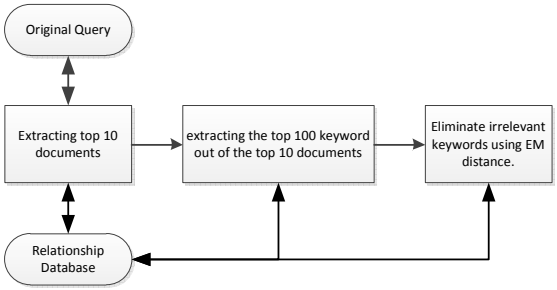

Fig. 2. Steps of Expanding Queries

For every input query, the system selects top 10 documents for each word of the query. Those documents are selected based on the importance of the query word to the entire text of the document. This approach was built on the assumption that the frequency of the word in a single document determines how important the word is to the subject after removing all stop words.

The importance of a word in a document is estimated based on the frequency of the word itself to the entire text of the document. As shown in equation (1).

$$IMP(X) = [1 - \frac{Freq(X)}{word\_Counter(D)}] * 100 \quad (1)$$

The importance of word X to document D is calculated as the percentage of appearance of word X in document D. The frequency of word X is divided by the total number of words in document D.

Next, the system selects the top 100 most important keywords used in these top 10 documents. The 100 words are selected using the same principle and equation as that used to select the top 10 documents. In other words, the system calculates the importance of each word in these documents and selects the top 100 words. EM distance is then used to eliminate keywords that are most likely not related to the original query that will be discussed in more details in the following section.

5 EM Algorithm

Expectation Maximization (EM) is a statistical method used for finding the maximum likelihood of parameters. EM is typically used to compute maximum likelihood estimates given incomplete samples. It is guaranteed to find a local optimum of data log likelihood. For the purpose of this research, EM is used to indicate similarity between two words based on their co-occurrence in a set of documents using equation (2).

$$EM(X,Y) = 1 - \frac{Freq\_documents(X,Y)}{Freq\_documents(X) + Freq\_documents(Y)} \quad (2)$$

EM distance between a word X and a word Y is calculated by dividing the total number of documents that both words appeared together by the sum of the total number of documents that each word appeared separately. In this situation, the EM distance indicates the degree to which word X and word Y are bonded, in terms of their concurrence in similar documents. We assume that the less the EM distance between two terms, the more bonded they are.

For an example if there are 10 documents that are talking about sports. Seven of which are talking about "كرة القدم" (football) while the rest are talking about "كرة السلة" (basketball). To simplify the example, we assume that each document that is talking about "كرة القدم" (football) has used this keywords once in its context and similarly the documents that are talking about "كرة السلة" (basketball). Calculating the EM distance between the word "كرة" (ball) and the word "القدم" (foot) using equation (3) will return:

$$\text{EM ("القدم", "كرة")} = 1 - \frac{7}{7+10} = 1 - 0.411 = 0.598 \quad (3)$$

Similarly calculating the distance between the word "كرة" (ball) and the word "السلة" (basket) using equation (4) will return:

$$\text{EM ("السلة", "كرة")} = 1 - \frac{3}{3+10} = 1 - 0.23 = 0.77 \quad (4)$$

You can see from the example above that the word "القدم" (foot) has shorter EM distance to the word "كرة" (ball) than when comparing to the EM distance between "كرة" (ball) and "السلة" (basket). Hence, the word "كرة" (ball) is more likely to come with the word "القدم" (foot).

An iterative approach is used to determine the best EM distance to describe the relevance between Arabic words. After experimentation, a threshold of 0.86 is reached as the optimal EM distance. The EM distances between the top 100 words and the query words are calculated, and all pairs of words with the EM distance more than 0.86 are considered not related to the query words. All other words that have an EM distance less than 0.86 are considered similar to the query words. These groups of words are used in the query expansion.

6 Evaluation

The evaluation used for this work is based on TREC evaluation procedure [6] that consists of a set of documents, a set of test topics and their relevance judgments. The INFILE corpus from CLEF 2009 initiative is selected as the evaluation corpus. Lu-

cene search engine was used for conducting the baseline search step with the original query. TRECEVAL software was used in calculating the precision, recall and other performance measures.

The INFILE corpus [17] from CLEF 2009 test consists of 50 different queries. Each query has a title, description and few keywords. The query titles consist of 2 to 6 words, while query description varies from few words to few lines. Keywords were not used in evaluating our approach.

We experimented with different parameters in the system in order to find the best approach. Table 1 shows the best runs for the baseline approach running the 50 queries. The baseline approach used only stemmed title words and was able to retrieve 1,007 of 1,195 the relevant documents. The second run used query description in its search and found 17 additional relevant documents. Combining title and description resulted in retrieving 1,078 relevant documents which has 71 extra relevant documents over retrieval with titles only. This shows that there are unique information about users' needs in both query titles and descriptions that complements each other. However, most studies on user behavior suggest that users rarely use long queries while searching on the web. Therefore run 2 and 3 are unlikely scenarios and the title search in run 1 is a better simulation of the user search behavior.

Table 1 shows that the precision decreased in the second run even though the number of query terms increased because of the length of the descriptions. This is due to the type of the terms found in the description itself and how tightly they are related to the meaning of the query. Query description might have the same meaning of the query, but it might not use as relevant keywords as those found in query title. Although the run with description only was able to pull more relevant documents, it was not able to put them in higher ranks. Moreover, when searching for both title and descriptions together, the system performance improved over both runs. The third run found more relevant documents and had higher precision in low recall which indicates better ranking. Unfortunately the queries in real life are closer to title only search rather than the second or third runs. This is because users do not write long descriptions for their searches. Many researchers such as Stefan Klink [8] have concluded that the average words used in a query is around two to three words which is very close to title only search in the first run in Table 1.

Run 1 (search for title only), Run 2 (search for description only), and Run 3 (search for title and description)

Many experiments were conducted to find the best configuration for query expansion. Table 2 lists two of the best runs with the proposed approach for the same queries. Run 4 shows the performance of title only search with query expansion. After expanding user query, the system was able to retrieve 12 more documents than the baseline and it has a better precision also (2.1% on recall 0). Run 4 used an EM distance equal to 0.86. All words that have EM distance less than 0.86 were considered similar to the query words. Run 5 is based on top of run 4. The only difference between run 4 and run 5 is that all terms in run 5 (documents and queries) have been stemmed. All the steps for processing queries and documents have been followed for run 5 on stemmed corpus. Run 5 is there to show that stemming [13] [14] didn't add

much value to the overall precision of the system. On the contrary it resulted in finding less relevant documents.

Table 1. Precision-Recall and the Precision at document cutoff for three baseline approaches

Recall	Run 1	Run 2	Run3
0	0.6182	0.6016	0.6374
0.1	0.5481	0.5094	0.5954
0.2	0.4763	0.4016	0.5309
0.3	0.4325	0.3535	0.5017
0.4	0.384	0.3078	0.4486
0.5	0.3652	0.2715	0.4141
0.6	0.3018	0.2195	0.3317
0.7	0.2734	0.1817	0.2896
0.8	0.2178	0.1490	0.2222
0.9	0.144	0.0930	0.1487
1	0.079	0.0553	0.0794

Doc. cut off	Run 1	Run 2	Run3
At 5 docs	0.44	0.3520	0.4600
At 10 docs	0.404	0.3420	0.4280
At 15 docs	0.3667	0.3240	0.3960
At 20 docs	0.331	0.3040	0.3740
At 30 docs	0.2873	0.2613	0.3307
At 100 docs	0.1554	0.1466	0.1714
At 200 docs	0.0889	0.0863	0.0971
At 500 docs	0.0384	0.0381	0.0414
At 1000 docs	0.0201	0.0205	0.0216

7 Conclusion

This research investigates query expansion using EM algorithm to improve the number of relevant documents retrieved. It also studies the best EM distance for Arabic words that describes the similarity between them. Results are compared with Lucene search results, which are used as a baseline. The test data used is the INFILE test corpus from CLEF 2009.

Moreover, the major contributions of this research are: 1) improving Arabic Information Retrieval through expanding Arabic queries with similar index terms, and 2) the list of suggested query terms and finding the best EM distance to define similarity between Arabic words.

Our experiments prove that expanding queries retrieves more relevant documents as shown in the evaluation section for queries than the baseline. Moreover, it also improves the overall recall precision for the final list of retrieved documents. The runs of the baseline show that there are unique information about users' information needs in both query titles and descriptions that complement each other. The experimental runs show that the proposed system is able to improve Arabic retrieval process while maintaining the same precision.

The EM distance is a major factor in the overall success of this system. It eliminates the unnecessary retrieved answers that the system was retrieving based on dissimilar keyword. It helps in focusing on those keywords that add value to the overall performance and query expansion and prevents the system from expanding the queries to dissimilar keywords. In future work we would like to experiment with EM algorithm focusing on text windows rather than whole documents for calculating the EM distance.

Table 2. Precision-Recall table and the Precision at document cutoff for best 2 trials of the Proposed Approach

Recall	Run 4	Run 5
0	0.6314	0.6381
0.1	0.5609	0.5645
0.2	0.4743	0.4819
0.3	0.4428	0.4430
0.4	0.3918	0.3948
0.5	0.3677	0.3694
0.6	0.3082	0.3063
0.7	0.2744	0.2818
0.8	0.2212	0.2194
0.9	0.1501	0.1451
1	0.0776	0.0778

Doc. cut off	Run 4	Run 5
At 5 docs	0.4560	0.4520
At 10 docs	0.4000	0.3900
At 15 docs	0.3693	0.3600
At 20 docs	0.3290	0.3300
At 30 docs	0.2887	0.2893
At 100 docs	0.1556	0.1544
At 200 docs	0.0886	0.0884
At 500 docs	0.0384	0.0385
At 1000 docs	0.0202	0.0202

8 Reference:

1. Staff, C. and Muscat, R.: Expanding Query Terms in Context. In: Proceedings of Computer Science Annual Workshop (CSAW'04), 106-108, University of Malta. (2004)
2. Bacchin, M., Melucci, M.: Expanding Queries using Stems and Symbols. In: Proceedings of the 13th Text REtrieval Conference (TREC 2004) Genomics Track, Gaithersburg, MD, USA, Nov (2004)
3. Martínez-Fernández, J., García-Serrano, A., Román, J., Paloma, M.: Expanding Queries Through Word Sense Disambiguation. In: 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006), Alicante, Spain, September 20-22, 2006, Revised Selected Papers. Carol Peters et al. (Eds.). LNCS, Springer, vol. 4730, pp. 613-616 (2007)
4. Manning, C.D., Raghavan, P., Schtze, H.: Relevance feedback and query expansion. In: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
5. Crestani, F.: Comparing neural and probabilistic relevance feedback in an interactive Information Retrieval system. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 3426-2430, Orlando, Florida, USA, June (1994)
6. http://trec.nist.gov/ (Visited on Sep 2010)
7. Magennis, M. van Rijsbcrgen, C.:The potential and actual effectiveness of interactive query expansion, In: Proceedings of ACM Special Interest Group in Information Retrieval Conference (SIGIR97), pp. 324-332 (1997)
8. Klink, S., Hust, A., Junker, M., Dengel, A.: Improving Document Retrieval by Automatic Query Expansion Using Collaborative Learning of Term-Based Concepts. Document Analysis Systems, pp. 376-387 (2002)
9. Khafajeh, H., Kanaan, G., Yaseen, M., Al-Sarayreh, B.: Automatic Query Expansion For Arabic Text Retrieval Based on Association and Similarity Thesaurus. In: Proceedings he European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS), Abu Dhabi, UAE (2010)
10. Rachidi, T., Bouzoubaa, M., ElMortaji, L., Boussouab, B., Bensaid, A.: Arabic user search Query correction and expansion. In: Proceedings of COPSTIC'03, Rabat December 11-13, (2003)

11. Bilotti, M.: Query expansion techniques for question answering. Master's thesis, Massachusetts Institute of Technology (2004)
12. Al-Shalabi, R., Kanaan, G., Yaseen, M., Al-Sarayreh, B., Al-Naji, N.: Arabic Query Expansion Using Interactive Word Sense Disambiguation. In: 2nd International Conference on Arabic Language Resources & Tools, MEDAR, April, pp. 156-158, Cairo, Egypt (2009)
13. Zitouni, A., Damankesh, A., Barakati, F., Atari, M., Watfa, M., Oroumchian, F.: Corpus based Arabic Stemming using N-grams, The Sixth Asia Information Retrieval Society Conference (AIRS2010), Taipei, Taiwan, LNCS, Springer, Vol. 6458, pp. 280-289 (2010)
14. Attia, M.: Arabic tokenization system. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 65-72 (2007)
15. Farghaly, A., Shaalan, K. Arabic Natural Language Processing: Challenges and Solutions, ACM Transactions on Asian Language Information Processing (TALIP), the Association for Computing Machinery (ACM), Vol. 8, 4:1-22 ACM Press (2009)
16. Nizar Y. Habash, Introduction to Arabic Natural Language Processing, (Synthesis lectures on human language technologies) Morgan & Claypool (2010)
17. Besançon, R. Chaudiron, S., Mostefa, D., Timimi, I. Choukri, K.: The INFILE Project: a Cross-lingual Filtering Systems Evaluation Campaign, In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, (2008).