A Framework for Automatic Query Expansion

Hazra Imran[1] and Aditi Sharan[2]

[1] Department of Computer Science, Jamia Hamdard, New Delhi, India
[2] School of Computers and System Sciences, Jawaharlal Nehru University, New Delhi, India
himran@jamiahamdard.ac.in, aditisharan@mail.jnu.ac.in

Abstract. The objective of this paper is to provide a framework and computational model for automatic query expansion using psuedo relevance feedback. We expect that our model can be helpful in dealing with many important aspects in automatic query expansion in an efficient way. We have performed experiments based on our model using TREC data set. Results are encouraging as they indicate improvement in retrieval efficiency after applying query expansion.

Keywords: Automatic Query Expansion (AQE), Pseudo Relevance Feedback (PRF), Information Retrieval (IR).

1 Introduction

In an information retrieval system, the query expansion is defined as an elaboration process of user's information need. Reformulation of the user queries is a common technique in information retrieval to cover the gap between the original user query and his need of information. Query expansion is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance. Efthimiadis [5] has done a complete review on the classical techniques of query expansion. Some of the important questions regarding query expansion s are: **What is the source of selecting expansion terms? Given the source, which terms should be selected for expansion? How should weights of terms be calculated?** Source of selecting the terms can be external or internal (from corpus itself). Considering corpus as source of selection, terms can be selected either globally (form entire corpus) or locally (from a subset of documents deemed to be relevant to query). Global analysis methods are computationally very expensive and their effectiveness is generally not better (sometimes worse) than local analysis. In local analysis user may be asked to select relevant documents from set of documents retrieved by information retrieval system. The problem with local analysis is that user's involvement makes it difficult to develop automatic methods for query expansion. To avoid this problem **pseudo relevance feedback (PRF)** approach is preferred in local analysis, where documents are retrieved using an efficient matching function and top n retrieved documents are assumed to be relevant. Automatic query expansion refers to techniques that modify a query without user control. One argument in favor of automatic query expansion is that the system has access to more statistical information on the relative utility of expansion terms and can make a better selection of which terms to add to the user's query. We have worked on local method for automatic query expansion using pseudo relevance feedback. In our previous work [7] we have

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 386–393, 2010.

focused on how thesaurus can be used for query expansion in selecting the terms externally.

In this paper we have proposed a framework for corpus based automatic query expansion. The paper is divided in V sections. In section II, we present a review of related work. Section III describes our proposed framework. Experimental results are presented in Section IV. Section V summarizes the main conclusions of this work.

2 Related Work

Early work of Maron [10] demonstrated the potential of term co-occurrence data for the identification of query term variants. Lesk[8] noted that query expansion led to the greatest improvement in performance, when the original query gave reasonable retrieval results, whereas, expansion was less effective when the original query had performed badly. Sparck Jones [14] has conducted the extended series of experiments on the ZOO-document subset of the Cranfield test collection. Sparck Jones results suggested that the expansion could improve the effectiveness of a best match searching.This improvement in performance was challenged by Minker et al. [11]. More recent work on query expansion has been based on probabilistic models [2]. Voorhees [4] expanded queries using a combination of synonyms, hypernyms and hyponyms manually selected from WordNet, and achieved limited improvement on short queries. Stairmand[9] used WordNet for query expansion, but they concluded that the improvement was restricted by the coverage of the WordNet. More recent studies focused on combining the information from both co-occurrence-based and handcrafted thesauri [13]. Carmel [3] measures the overlap of retrieved documents between using the individual term and the full query. Ruch et al.[12] studied the problem in the domain of biology literature and proposed an argumentative feedback approach, where expanded terms are selected from only sentences classified into one of four disjunct argumentative categories. Cao [6] uses a supervised learning method for selecting good expansion terms from a number of candidate terms.

3 Proposed Framework

Improving robustness of query expansion has been goal of researchers in last few years. We propose a framework for automatic query expansion using Pseudo Relevance feedback. Our framework allows to experiment on various parameters for automatic query expansion. Figure 1 depicts the components in our proposed framework. For a given query, *Information Retrieval system* will fetch top N documents from the corpus similar to the query, based on some similarity measures such as jaccard and okapi similarity measure. *Summary Generation System* takes the ranked Top N documents as input and generates the summary corresponding to each document. Either top N documents or their corresponding summaries will act as *Source for Selecting Expansion Terms*.

Once the source is selected, our next module is for extracting expansion terms. We have used two *Methods for extracting terms*: - one is based on Term co-occurrences and other is based on Lexical Links. Based on term co-occurrence method, we derive expansion terms that are statistically co-occurring with the given query. For our

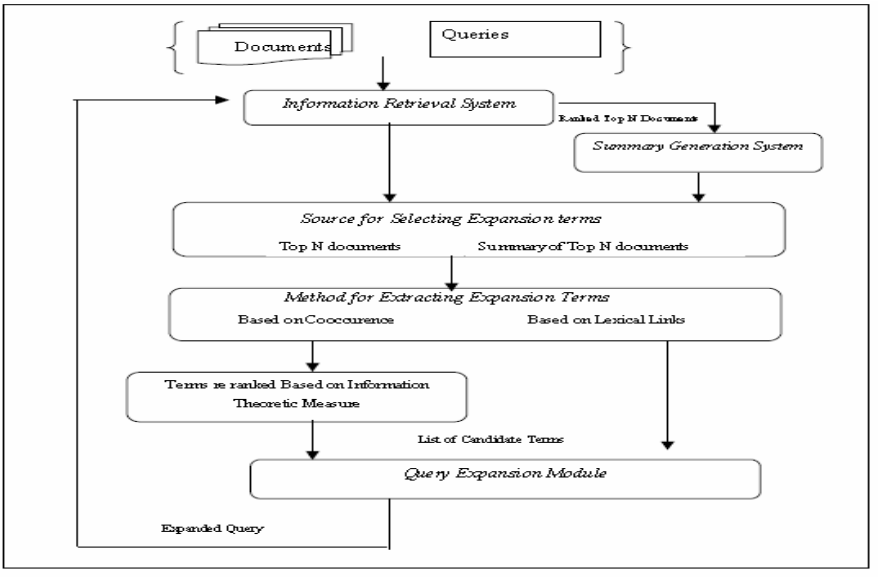

Fig. 1. Proposed automatic query expansion and retrieval framework

experiments, we have used jaccard coefficient for measure the similarity between query terms and all other terms present in relevant documents.

$$jaccard\_co(t_i, t_j) = \frac{d_{ij}}{d_i + d_j - d_{ij}} \tag{1}$$

Where

d_i and d_j are the number of documents in which terms t_i and t_j occur, respectively, and d_{ij} is the number of documents in which t_i and t_j co-occur.

However there is a danger in adding these terms directly the query. The candidate terms selected for expansion could co-occur with the original query terms in the documents (top n relevant) by chance. The higher its degree is in whole corpus, the more likely it is that candidate term co-occurs with query terms by chance. Keeping this factor in mind inverse document frequency of a term can be used along with above discussed similarity measures to scale down the effect of chance factor. Incorporating inverse document frequency and applying normalization define degree of co-occurrence of a candidate term with a query term as follows:

$$co\_degree(c, t_j) = \log_{10}(co(c, t_j) + 1) * (idf(c) / \log_{10}(D)) \tag{2}$$

$$idf(c) = \log_{10}(N / N_c) \tag{3}$$

Where

N = number of documents in the corpus
D= number of top ranked documents used
c =candidate term listed for query expansion

N_c = number of documents in the corpus that contain c

$co(c,t_j)$ = number of co-occurrences between c and t_j in the top ranked documents i.e. $jaccard\_co(c_i,t_j)$

To obtain a value measuring how good c is for whole query Q, we need to combine its degrees of co-occurrence with all individual original query terms. So we use suitability for Q to compute $t_1, t_2, \ldots t_n$

$$Suitabiltyfor Q = f(c,Q) = \prod_{t_i in Q} (\delta + co\_deg ree(c,t_i))^{idf(t_i)} \quad (4)$$

Above equation provides a suitability score for ranking the terms co-occurring with entire query. Still there are chances that a term that is frequent in top n relevant documents is also frequent in entire collection. In fact this term is not a good for expansion, as it will not allow discriminating between relevant and non-relevant document. Keeping this as motivation we suggest the use of information theoretic measures for selecting good expansion terms. We then rank the expansion terms based on the KLD. This approach is based on studying the difference between the term distribution in the whole collection and in the subsets of documents that are relevant to the query, in order to, discriminate between good expansion terms and poor expansion term. Terms closely related to those of the original query are expected to be more frequent in the top ranked set of documents retrieved with the original query than in other subsets of the collection or entire collection. We used the concept of Kullback-Liebler Divergence to compute the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained using the original user query. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection. For the term t this divergence is:

$$KLD(t) = [p_R(t) - p_C(t)] \log \frac{p_R(t)}{p_C(t)} \quad (5)$$

where $P_R(t)$ be the probability of t estimated from the corpus R, $P_C(t)$ is the probability of $t \in V$ estimated using the whole collection. To estimate $P_C(t)$, we used the ratio between the frequency of t in C and the number of terms in C.

$$P_C(t) = \frac{f(t)}{N} \quad (6)$$

$$P_R(t) = \begin{cases} \gamma \frac{f_R(t)}{NR} if t \in v(R) \\ \delta p_c(t) otherwise \end{cases} \quad (7)$$

Where C is the set of all documents in the collection, R is the set of top retrieved documents relative to a query, *NR* is the number of terms in R ,v(R) be the vocabulary of all the terms in R, $f_R(t)$ is the frequency of t in R ,f(t)=frequency of term in C,N is the number of terms in C.

The candidate terms were ranked by using equation (7) with γ=1, which amounts to restricting the candidate set to the terms contained in R. Terms not present in relevant

sent are given a default probability. The other method is based on Lexical links. The method relies on calculating lexical cohesion between query terms' contexts in a document [15]. The main goal of this method is to rank documents by taking into consideration how cohesive the contextual environments of distinct query terms are in each document. The assumption is that if there is a high degree of lexical cohesion between the contexts of distinct query terms in a document, they are likely to be topically related, and there is a greater chance that this document is relevant to the user's query. Finally *query expansion module* reformulates the query by adding potential candidate terms with the initial query. We have given the weights to expanded query terms using their tf X idf values. The document collection is then ranked against the reformulated query.

4 Experiments

We have used the Vector Space Model implementation to build our information retrieval system. Stemming and stop word removing has been applied in indexing and expansion process. For our experiments, we used volume 1 of the *TIPSTER* document collection, a standard test collection in the IR community. We have used WSJ corpus, and TREC topic set, with 50 topics, of which we only used the title (of 2.3 average word length) for formulating the query. We have used different measures to evaluate each method. The measures considered have been MAP (Mean Average Precision), Precision@5, and Precision @10.

Parameters for performing Automatic Query Expansion using Pseudo Relevance Feedback

Firstly, we investigate the parameters for performing AQE having effect on retrieval performance. The parameters of query expansion are: *Top N doc*←number of top-ranked documents to be considered as the pseudo-relevance set), *Number of expansion terms*←the number of informative terms to be added to the query).

1. Number of top ranked documents

Based on the fact that the density of relevant documents is higher for the top-ranked documents, one might think that the fewer the number of documents considered for expansion, the better the retrieval performance. However, this was not the case. As shown in Table1, the retrieval performance was found to increase as the number of documents increased, at least for a small number of documents, and then it gradually dropped as more documents were selected.

This behavior can be explained considering that the percentage of truly relevant documents in the pseudo-relevant documents is not the only factor affecting

Table 1. Performance versus number of pseudo-relevant documents for TREC-1

	5	10	15	20	25	30
Mean Average Precision	0.129902	0.129906	0.129910	0.129696	0.1295	0.12924
PREC-AT-5	0.1344333	0.1342334	0.1344334	0.1344334	0.13243	0.13243
PREC-AT-10	0.1201	0.1202	0.1211698	0.12018	0.12019	0.1191698

performance here. The optimal choice should represent a compromise between the maximization of the percentage of relevant documents and the presence of at least some relevant document. Consistently with the results reported in Table 1, we found that these two parameters were best balanced when the size of the training set ranged from 6 to 15. Further experiments revealed that the system performance decreased nearly monotonically as the number of documents was increased beyond those shown in Table 1. The decline in performance was however slow, because the number of relevant documents remained substantially high even after a large number of retrieved documents. For instance, for TREC-1, the average precision at 20 documents was 0.129696, at 30 documents was 0.12924.

2. Number of terms selected for expansion

This section seeks to answer questions regarding the best number of query terms to use for expansion. We let the number of expansion terms vary from 5 to 30 (step = 5), computing for each value the retrieval performance of the system. Table 2 shows that the maximum values of the different performance measure were reached for different choices of the number of selected terms. Most important, the results show that the variations in performance were negligible for all measures and for all selected sets of expansion terms. With respect to *number of expansion terms* considered in the QE, using less than 10 terms means a drop-off in MAP, while for *number of expansion terms* $\geq$ 10, the retrieval performance is stable. To assess the stability of the approaches with respect to *number of* expansion term and top N doc, we vary them and record the MAP. In particular, we vary 2 $\leq$ top N doc $\leq$ 30 and 1 $\leq$ number of expansion term $\leq$ 20.

Table 2. Performance versus number expansion terms for TREC-1

	5	10	15	20
Mean Average Precision	0.129906	0.132393	0.139817	0.139802
PREC-AT-5	.1330334	0.1342234	0.1344334	0.1341334
PREC-AT-10	0.1172	0.1200169	0.1211698	0.1211698

Figure 2 present surface plots of Query Expansion Settings for KLD.

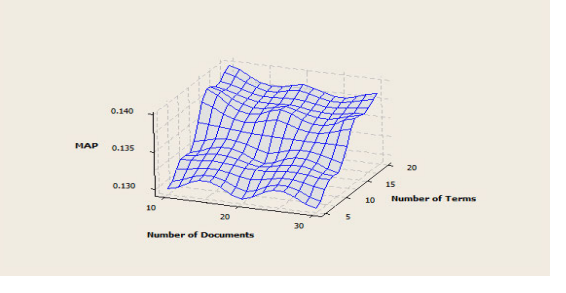

Fig. 2. Surface plots of MAP for Query Expansion when the number of Top N documents and number of Expansion Terms parameters are varied

In our framework, we have used two sources for selecting expansion terms. One is a top N document and other is summary of top N documents. Again, we suggest two methods for extracting terms .One is based on co-occurrence and other is on lexical links. Figure 3 shows the graph of an example of a query where terms are extracted based on term co-occurrence , KLD and lexical information respectively on different queries.

Analysis of result shows that after intensive computation we may select appropriate parameters for AQE . Further, we observe that re-ranking co-occurring terms on KLD, improve result significantly. For method based on lexical link we observe that sufficient links are available only for few queries. However, for the queries where sufficient links are found, retrieval performances improved in most of the cases.

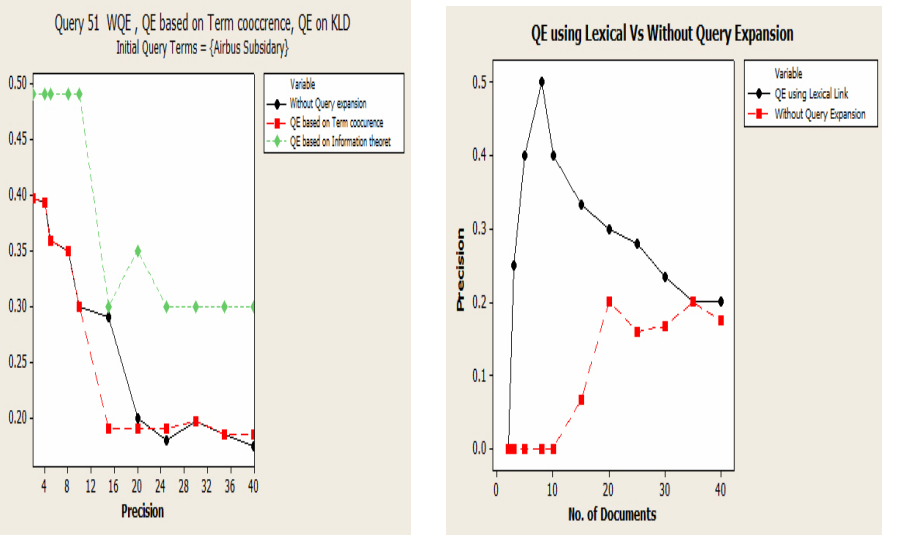

Fig. 3. Graph of a particular query for without query expansion, QE based on Term co-occurrence and information theoretic measure and Lexical Information

5 Conclusion

In this paper we have proposed a framework along with a computational model for automatic query expansion using PRF. Our framework is flexible and feasible. Regarding flexibility, it allows you to experiment with different methods for selecting top n relevant, selecting query expansion terms, and selecting parameters for query expansion. Regarding feasibility it provides step-by-step procedure for implementing query expansion. Analysis of our results shows that query terms selected on the basis of co-occurrence are related to original query, but may not be good discriminator to discriminate between relevant and non-relevant document. KLD measure allows this discrimination to certain extent; hence it improves retrieval performance over co-occurrence based measure. Lexical links allows us to deal with the context of query in addition to co-occurrence measure. We are exploring use of semantic links for improving lexical based query expansion.

References

1. Lee, C.J., Lin, Y.C., Chen, R.C., Cheng, P.J.: Selecting effective terms for query formulation. In: Proc. of the Fifth Asia Information Retrieval Symposium (2009)
2. Croft, W.B., Harper, D.J.: Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35, 285–295 (1979)
3. Carmel, D., Yom-Tov, E., Soboroff, I.: SIGIR Workshop Report: Predicting query difficulty – methods and applications. In: Proc. of the ACM SIGIR 2005 Workshop on Predicting Query Difficulty – Methods and Applications, pp. 25–28 (2005)
4. Voorhees, E.M.: Query expansion using lexical semantic relations. In: Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval (1994)
5. Efthimiadis, E.N.: Query expansion. Annual Review of Information Systems and Technology 31, 121–187 (1996)
6. Cao, G., Nie, J.Y., Gao, J.F., Robertson, S.: Selecting good expansion terms for pseudorelevance feedback. In: Proc. of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250 (2008)
7. Imran, H., Sharan, A.: Thesaurus and Query Expansion. International journal of computer science & information Technology (IJCSIT) 1(2), 89–97 (2009)
8. Lesk, M.E.: Word-word associations in document retrieval systems. American Documentation 20, 27–38 (1969)
9. Stairmand, M.A.: Textual context analysis for information retrieval. In: Proceedings of the 1997 ACM SIGIR Conference on Research and Development in Information Retrieval (1997)
10. Maron, M.E., Kuhns, J.K.: On relevance, probabilistic indexing and information retrieval. Journal of rhe ACM 7, 216–244 (1960)
11. Minker, J., Wilson, G.A., Zimmerman, B.H.: Query expansion by the addition of clustered terms for a document retrieval system. Information Storage and Retrieval 8, 329–348 (1972)
12. Ruch, P., Tbahriti, I., Gobeill, J., Aronson, A.R.: Argumentative feedback: A linguistically-motivated term expansion for information retrieval. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 675–682 (2006)
13. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval (1999)
14. Sparck Jones, K.: Automatic keyword classification for information retrieval. Butterworth, London (1971)
15. Vechtomova, O., Wang, Y.: A study of the effect of term proximity on query expansion. Journal of Information Science 32(4), 324–333 (2006)