

SELECTING EFFECTIVE EXPANSION TERMS FOR BETTER INFORMATION RETRIEVAL

Hazra Imran

Department of Computer Science
Jamia Hamdard , NewDelhi, India
himran@jamiahamdard.ac.in

Aditi Sharan

School of Computers and System Sciences
Jawaharlal Nehru University, New Delhi, India
aditisharan@mail.jnu.ac.in

Automatic Query expansion is a well-known method to improve the performance of information retrieval systems. In this paper, we consider methods to extract the candidate terms for automatic query expansion, based on co-occurrence information from psuedo relevant documents. The objective of the paper is: to present to user different ways of selecting and ranking co-occurring terms and to suggest use of information theoretic measures for ranking the co-occurring terms selected, in order to improve retrieval efficiency. Specifically in our work, we have used two information theoretic measures: *Kullback-Leibler* divergence (KLD) and a variant of KLD. These measures are based on relative entropy between top documents and entire collection. We have compared the retrieval improvement achieved by expanding the query with terms obtained with different methods belonging to both approaches (co occurrence based and information theoretic). Experiments have been performed on TREC-1 data set. Intensive experiments have been done to select suitable parameters used in automatic query expansion, such as number of top n selected documents and the number of terms selected for expansion. Results suggest that firstly considerable improvements can be achieved if co-occurring terms are selected properly by considering different options available for selecting them. Secondly, information theoretic measures applied over co-occurring terms can be helpful in improving retrieval efficiency.

Keywords: Automatic Query Expansion, Candidate Terms, Term Co-occurrence, Kullback-Leibler divergence, Pseudo Relevance Feedback

1. Introduction

Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance. Current information retrieval systems are limited by many factors reflecting the difficulty to satisfy user requirements expressed by short queries. 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. The most used technique for query reformulation is query expansion, where the original user query is expanded with new terms extracted from different sources. Efthimiadis [7] has done a complete review on the classical techniques of query expansion. The main problem of query expansion is that in some cases the expansion process worsens the query performance. Improving the robustness of query expansion has been the goal of many researchers in the last years.

The terms used for expansion can be selected from external sources or from the corpus itself. In our previous work, [12] we have focused on how a thesaurus can be used for query expansion in selecting the terms externally. In this paper we have used the corpus to select query expansion terms. Some of the methods for selecting the terms from the corpus are based on global analysis, where the list of candidate terms is generated from the whole collection, where as others are based on local analysis [32,15,16] where the expanding terms are selected by user.

Global analysis methods are computationally very expensive and their effectiveness is generally not better (sometimes worse) than local analysis. The problem with local analysis is that user's feedback is required to provide information regarding top rank relevant documents. User's involvement makes it difficult to develop automatic methods for query expansion. To avoid this problem pseudo relevant feedback approach is preferred where documents are retrieved using an efficient matching function and top n retrieved documents are assumed to be relevant. Value of n has to be selected empirically.

In this paper we have performed local query expansion based on pseudo- relevance feedback. In this work, we have tested different approaches to extract the terms co-occurring with original query and selected best approach based on theoretical as well as empirical observations. Further we have suggested use of information theoretic measures for ranking the co-occurring terms selected, in order to improve retrieval efficiency. Specifically in our work, we have used two information theoretic measures: *Kullback-Leibler* divergence (KLD) and a variant of KLD. These measures are based on relative entropy between top documents and entire collection. Experiments have been performed on TREC-1 data set. Intensive experiments have been done to select suitable parameters used in automatic query expansion, such as number of top n selected documents and the number of terms selected for expansion. The paper is divided in sections. In section 2, we present a review of related work. Sections 3 and 4 describe the co-occurrence and information-theoretic approaches, respectively; Section 5 describes our methodology. The experimental results are presented in Section 6 and Section 7 summarizes the main conclusions of this work.

2. Related Work

Early work of Maron [21] demonstrated the potential of term co-occurrence data for the identification of query term variants. Lesk [18] expanded a query by the inclusion of terms that had a similarity with a query term greater than some threshold value of the cosine coefficient. Lesk 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 [30] has conducted the extended series of experiments on the ZOO-document subset of the Cranfield test collection. The terms in this collection were clustered using a range of different techniques and the resulting classifications were then used for query expansion. Sparck Jones results suggested that the expansion could improve the effectiveness of a best match searching, if only, the less frequent terms in the collection were clustered with the frequent terms being unclustered and if only, very similar terms were clustered together. This improvement in performance was challenged by Minker et al.[22]. More recent work on query expansion has been based on probabilistic models of the retrieval process and has tried to relax some of the strong assumptions of a term statistical independence that normally needs to be invoked, if probabilistic retrieval models are to be used [4,26]. In a series of papers, Van Rijsbergen had advocated the use of query expansion techniques based on a minimal spanning tree (MST), which contains the most important of the inter-term similarities calculated using the term co-occurrence data and which is used for expansion by adding in those terms that are directly linked to query terms in the MST [13,29,2,31]. Later work compared relevance feedback using both expanded and nonexpanding queries and using both MST and non-MST methods for query expansion on the Vaswani test collection [28,29]. Voorhees [6] expanded queries using a combination of synonyms, hypernyms and hyponyms manually selected from WordNet, and achieved limited improvement on short queries. Stairmand [19] used WordNet for query expansion, but they concluded that the improvement was restricted by the coverage of the WordNet and no empirical results were reported. More recent studies focused on combining the information from both co-occurrence-based and handcrafted thesauri [24,25]. Liu et al.[27] used WordNet for both sense disambiguation and query expansion and achieved reasonable performance improvement. However, the computational cost is high and the benefit of query expansion using only WordNet is unclear. Carmel [5] measures the overlap of retrieved documents between using the individual term and the full query. Previous work [1] attempt to sort query terms according to the effectiveness based on a greedy local optimum solution. Ruch et al.[23] 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 [11] uses a supervised learning method for selecting good expansion terms from a number of candidate terms.

3. Co-occurrence Based Approach for selecting terms for query expansion

The methods based on the term co-occurrence which have been used since the 70's to identify the relationships that exist among terms. Van Rijsbergen [2] has given the idea of using co-occurrence statistics to detect the semantic similarity between terms and exploiting it to expand the user's queries.

In this approach the terms co-occurring with original query terms are selected as candidate terms for expansion. In fact, the idea is based on the Association Hypothesis:

“If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is likely to be good at this.”

Main source of extracting co-occurring terms is the corpus from where documents are coming. As discussed in previous section these terms can be selected globally or locally, each having their pros and cons. Further for local selection top n documents can be selected based on users feedback or psuedo-relevance feedback information. The main problem with the co-occurrence approach was mentioned by Peat and Willet [14] who claims that similar terms identified by co-occurrence tend to occur also very frequently in the collection and therefore, these terms are not good elements to be discriminate between relevant and non-relevant documents. This is true when the co-occurrence analysis is done generally on the whole collection but if we, apply it only on the top ranked documents discrimination does occur to a certain extent.

The basic question with this approach is how to select co-occurring terms, as terms can be selected in a number of ways. However some standard measures have been suggested to select co-occurring terms.

For our experiments, we have used two well-know coefficients: - jaccard and frequency

$$jaccard_co(t_i, t_j) = \frac{d_{ij}}{d_i + d_j - d_{ij}} \quad (1)$$

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.

$$freq_co(t_i, t_j) = \sum_{d \in D} (f_{d,t_i} \times f_{d,t_j}) \quad (2)$$

t_i = all terms of top N docs terms

t_j = query term

f_{d,t_i} = frequency of term t_i in doc d
 f_{d,t_j} = frequency of term t_j in doc
 D = number of top ranked documents used

We can apply these coefficients to measure the similarity between terms represented by the vectors. However there is a danger in adding these terms directly to 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:

$$co_degree(c, t_j) = \log_{10}(co(c, t_j) + 1) * (idf(c) / \log_{10}(D)) \quad (3)$$

$$idf(c) = \log_{10}(N / N_c) \quad (4)$$

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) or freq_co(c_i, t_j).

Above formula can be used for finding similarity of a term c with one query term. 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 $t_1, t_2 \dots t_n$. So we use suitabilityforQ to compute

$$SuitabilityforQ = f(c, Q) = \prod_{t_i \in Q} (\delta + co_degree(c, t_i))^{idf(t_i)} \quad (5)$$

Above equation provides a suitability score for ranking the terms co-occurring with entire query.

Now we can summarize the co-occurrence approaches used by us in this work. We can use either jaccard_coefficient (eqn 1) or freq_coefficient (eqn 2) measure to select the

terms co-occurring with query terms. Suitability formula (eqn 5) provides an efficient way to combine all terms co-occurring with individual query terms.

4. Information-Theoretic Approach as a tool for ranking co-occurring terms

4.1. Kullback-Liebler Divergence Measure

Now we again refer to Peat and Willet [14] who claim that similar terms identified by co-occurrence tend to occur also very frequently in the collection and therefore, these terms are not good elements to be discriminate between relevant and non-relevant documents. If the co-occurring terms are selected from top ranked documents discrimination does occur to a certain extent. However 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. 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.

One of the most interesting approaches based on term distribution analysis has been proposed by Claudio et al [3], who uses the concept the 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:

$$KLD(t) = [p_R(t) - p_C(t)] \log \frac{p_R(t)}{p_C(t)} \quad (6)$$

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.

$$P_C(t) = \frac{f(t)}{N} \quad (7)$$

$$P_R(t) = \begin{cases} \gamma \frac{f_R(t)}{NR} & \text{if } t \in v(R) \\ \delta p_c(t) & \text{otherwise} \end{cases} \quad (8)$$

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

The candidate terms were ranked by using equation (6) with $\gamma=1$, which amounts to restricting the candidate set to the terms contained in R.

4.2. Kullback- Liebler Divergence Variation Measure

The KLD measure suggested in previous subsection treats all the terms equally. In other words it does not consider importance of term based on the ranked order of document, from where term is coming. Therefore we suggest use of a variant of KLD, which takes into account the likely degree of relevance of the documents retrieved in the initial run, in order to determine importance of a term. In this variation $f(t)/NR$ is replaced by another function giving equation 9.

$$KLD_variation(t) = [p_R(t) - p_C(t)] \log \frac{P_R(t)}{p_C(t)} \quad (9)$$

$$P_R(t) = \frac{\sum_d f(t) \times score_d}{\sum_t \sum_d f(t) \times score_d} \quad (10)$$

5. Our methodology

We have performed local query expansion based on pseudo relevance feedback. Following are the steps in our methodology.

1. *Identify initial set of tokens* -- Our system first identified the individual terms occurring in the document collection.
2. *Preprocessing* -- Stop word removal and stemming is performed. We used porter-stemming algorithm [20].
3. *Document weighting*. We assigned weights to the terms in each document by the classical *tf.idf* scheme.
4. *Weighting of unexpanded query*: To weigh terms in unexpanded query, we used the *tf* scheme.
5. *Document ranking with unexpanded query*: We computed a document ranking using common coefficients jaccard between the document vectors and the unexpanded query vector.
6. *Select top n documents for extracting candidate terms*.
7. *Listing of candidate terms*: We use *jacc_coefficient* or *freq_coefficient* using equation (1) or (2) to list out the candidate terms from top n documents. These candidate terms can be used for expansion.
8. *Expansion term ranking using suitability value*: The candidate terms were ranked using equation (5).
9. *Expansion term ranking using information theoretic measure*: Re-rank the terms obtained in step 8 using equation (6)(KLD) or (9) (KLD variant).
10. *Construction of expanded query*: We added the ranked terms to the original query.
11. *Document ranking with expanded query*: The final document ranking was computed by using jaccard coefficient between the document vectors and the expanded query vector.
12. *Compare performance of results for unexpanded and expanded query using precision recall based measures*.

6. Experiments

For our experiments, we used volume 1 of the *TIPSTER* document collection, a standard test collection in the IR community. Volume 1 is a 1.2 GByte collection of full-text articles and abstracts. The documents came from the following sources.

WSJ -- Wall Street Journal (1986, 1987, 1988, 1989,1990,1991 and 1992)
AP -- AP Newswire (1988,1989 and 1990)
ZIFF -- Information from Computer Select disks (Ziff-Davis Publishing)
FR -- Federal Register (1988)
DOE -- Short abstracts from Department of Energy

We have used WSJ (1991-1992) corpus, and TREC topic set, with 50 topics, of which we only used the title (of 2.3 average word length). In our first approach, suitability formula

was used for selecting the expansion terms (equation (5) (jaccard_coefficient is used to select the similar terms). These terms were then ranked using KLD measure (equation (6)). In a similar way, for the third approach we used a variant of KLD(equation 9) in order to select the subset of terms from the terms selected by suitability value. We have compared the result of all these approaches with that of unexpanded query.

We have used different measures to evaluate each method. The measures considered are MAP (Mean Average Precision), Precision@5, Precision @10. Precision and recall are general measures to quantify overall efficiency of a retrieval system. However, when a large number of relevant documents are retrieved overall precision and recall values do not judge quality of the result. A retrieval method is considered to be efficient if it has high precision at low recalls. In order to quantify this precision can be calculated at different recall levels. We have calculated Precision@5, Precision@10 recall level.

Parameter Study

We have studied two parameters that are fundamental in query expansion: number of candidate terms to expand the query and number of top N documents used to extract the candidate terms. The optimal value of these parameters can be different for each method, and thus we have studied them for each case. Following graphs shows the result for different parameter values for each of the methods.

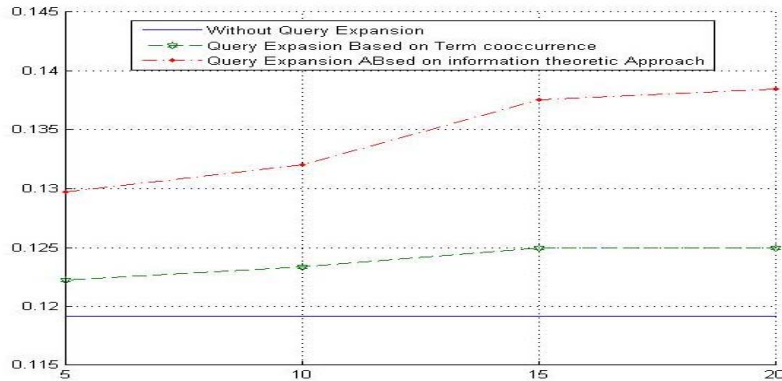


Fig. 1. Curve showing the MAP with different numbers of candidate terms to expand the original query using Query Expansion based on co-occurrence and information theoretic measure.

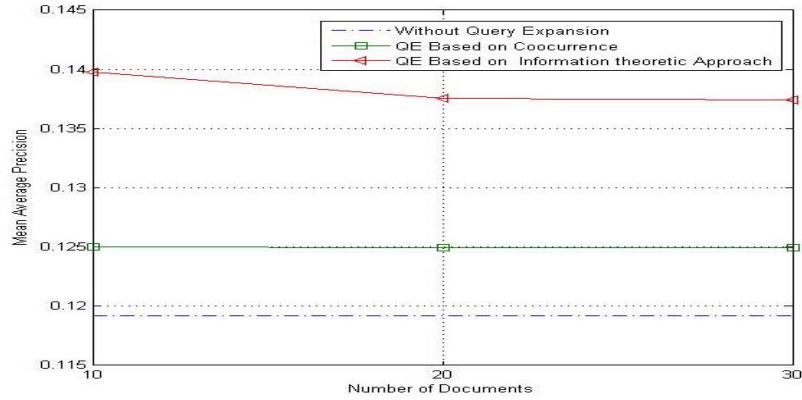


Fig. 2. Curve showing the MAP measures with different numbers of top documents used to extract the set of candidate query terms.

We can observe that in all cases the best value for number of document selected for query expansion is around 20 documents and for the number of query expansion terms is 15. This implies that there is a certain threshold on number of documents and number of query expansion terms to be added in order to improve efficiency of query expansion.

Comparative Analysis of Result

Table 1 shows overall comparative result for all query expansion methods considered in our work. The parameter values for number of top documents is 20 and number of query terms to be added are 15. From the table we can observe that in general, terms selected with KLD are better than suitability ranking. We also observed that with the KLD we are able to improve the overall precision (MAP). In some cases, KLD_variant is able to improve precision@5. By changing various parameters, we may be able to visualize the effect of KLD_variant.

Table 1. Comparative result for query expansion methods used in our work.

	MAP	P@5	P@10
Unexpanded query approach	0.119176	0.1224269	0.120552
Candidate term ranking using Suitabilty of Q with jaccard coefficient	0.124961	0.12482	0.1317
Candidate term ranking using Suitabilty of Q with frequency coefficient	0.120272	0.1194349	0.12062
Candidate term ranking using KLD with jaccard coefficient	0.137432	0.132433	0.11917
Candidate term ranking using KLD with frequency coefficient	0.126466	0.131773	0.11704
KLD_variation with jaccard coefficient	0.12848	0.13577	0.11544
KLD_variation with frequency coefficient	0.12531	0.132773	0.11686

7. Conclusions and Future Works

In this paper we have discussed different approaches of improving retrieval efficiency by carefully selecting co-occurring terms used for automatic query expansion. Firstly, present to user different ways of selecting and ranking co-occurring terms. Further, we suggest use of information theoretic measures for ranking the co-occurring terms selected, in order to improve retrieval efficiency. Specifically in our work, we have used two information theoretic measures: *Kullback-Leibler* divergence (KLD) and a variant of KLD. These measures are based on relative entropy between top documents and entire collection. Experiments have been performed on TREC-1 data set. Results suggest that considerable improvements can be achieved if co-occurring terms are selected properly by considering different options available. We also observed that information theoretic measures applied over co-occurring terms could be helpful in improving retrieval efficiency. We have used co-occurrence measures for selecting the terms but other criteria, such as the coherence between the candidate terms, can also be useful in selecting better candidate terms.

References

1. C. J. Lee, Y. C. Lin, R. C. Chen, and P. J. Cheng. Selecting effective terms for query formulation. In *Proc. of the Fifth Asia Information Retrieval Symposium*, 2009.

2. C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, (33): pages 106–119, 1977.
3. Carpineto, C. and Romano, G. 2000b. TREC-8 automatic ad-hoc experiments at FUB. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)* (Gaithersburg, Md.), NIST Special Publication 377–380,1999.
4. Croft, W. B., & Harper, D. J. ,Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35, pages 285-295,1979.
5. D. Carmel, E. Yom-Tov, and I. Soboroff. SIGIR Workshop Report: Predicting Query Difficulty -- methods and applications. In *Proc. of the ACM SIGIR 2005 Workshop on Predicting Query Difficulty -- Methods and Applications*, pages 25--28, 2005.
6. E. M. Voorhees. Query expansion using lexical semantic relations. In *Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval, 1994*.
7. E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.
8. Ellen M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *TREC*,pages 69–77, 2003.
9. Ellen M. Voorhees. The TREC 2005 robust track. *SIGIR Forum*, 40(1): 41–48, 2006.
10. Ellen M. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39(1): pages 11–20,2005.
11. G. Cao, J. Y. Nie, J. F. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243--250, 2008.
12. H.Imran, A.Sharan.” Thesaurus and Query Expansion”, *International journal of computer science & information Technology (IJCSIT)*, Vol 1, No 2, pages 89-97,2009.
13. Harper, D. J., & van Rijsbergen, C. J. Evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34, pages 189-216,1978.
14. Helen J. Peat and Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS*, 42(5): pages 378–383, 1991.
15. Hinrich Schütze and Jan O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Inf. Process. Manage*, 33(3): pages 307–318, 1997.
16. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference*, pages 146–160, New York, US, 1994.
17. Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1): pages 79–112, 2000.
18. Lesk, M. E. . Word-word associations in document retrieval systems. *American Documentation*, 20, pages 27-38,1969.
19. M. A. Stairmand. Textual context analysis for information retrieval. In *Proceedings of the 1997 ACM SIGIR Conference on Research and Development in Information Retrieval*,1997.
20. M.F. Porter. An algorithm for suffix stripping, in *Program - automated library and information systems*, 14(3): pages 130-137,1980.
21. Maron, M. E., & Kuhns, J. K. . On relevance, probabilistic indexing and information retrieval. *Journal of rhe ACM*, 7 , pages 216- 244,1960.
22. 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, pages 329-348 ,1972
23. P. Ruch, I. Tbahriti, J. Gobeill, and A. R. Aronson. Argumentative feedback: A linguistically-motivated term expansion for information retrieval. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 675–682,2006.

24. R. Mandala, T. Tokunaga, and H. Tanaka. 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.
25. R. Mandala, T. Tokunaga, and H. Tanaka. Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. In *Proceedings of the seventh Text REtrieval Conference (TREC7)*, 1999.
26. Robertson, S. E., & Sparck Jones, K. Relevance weighting of search terms. *Journal of the American Society of Informarion Science*, 21, pages 129-146, 1976.
27. S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.
28. Smeaton, A. F. *The retrieval effects of query expansion on a feedback document retrieval system*, University College Dublin, MSc thesis, 1982.
29. Smeaton, A. F., & van Rijsbergen, C. J. The retrieval effects of query expansion on a feedback document retrieval system. *Computer Journal*, 26, pages 239-246, 1983.
30. Sparck Jones, K. *Automatic keyword classification for information retrieval*, London: Butterworth , 1971.
31. Van Rijsbergen, C. J., Harper, D. J., & Porter, M. F. The selection of good search terms. *Information Processing and Management*, 17, pages 77-91, 1981.
32. Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In SIGIR, pages 160-169, 1993.