Improving Effectiveness of Query Expansion Using Information Theoretic Approach

Hazra Imran[1] and Aditi Sharan[2]

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

Abstract. Automatic Query expansion is a well-known method to improve the performance of information retrieval systems. In this paper we have suggested information theoretic measures to improve efficiency of co-occurrence based automatic query expansion. We have used pseudo relevance feedback based local approach. The expansion terms were selected from the top N documents using co-occurrence based approach. They were then ranked using two different information theoretic approaches. First one is standard *Kullback-Leibler* divergence (KLD). As a second measure we have suggested use of a variant KLD. Experiments were performed on TREC-1 dataset. The result suggests that there is a scope of improving co-occurrence based query expansion by using information theoretic measures. Extensive experiments were done to select two important parameters: number of top N documents to be used and number of terms to be used for expansion.

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

1 Introduction

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 widely used technique for query reformulation is query expansion, where the original user query is expanded with new terms extracted from different sources. Queries submitted by users are usually very short. 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 and most proposed approaches use external collections [8,9,10] to extract candidate terms for the expansion. In our previous work, [12] we have focused on how a thesaurus can be used for query expansion.

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part II, LNAI 6097, pp. 1–11, 2010.

Query Expansion can be: Manual, semiautomatic and automatic. In corpus- based automatic query expansion the terms to be added to the query can either be selected globally (from the entire document collection) or locally (from top N retrieved documents). Methods based on global analysis are computationally very expensive and its effectiveness is not better than that of methods based on local analysis [32,15,16]. Xu and Croft [17] have suggested the use of local context analysis (LCA) to achieve tradeoff between local and global query expansion. Our work relates to automatic query expansion done locally.

Most of the automatic query expansion methods use co-occurrence based approach to select the terms for query expansion. However, this is very broad and general approach and all the co-occurring terms don't have equal probability of improving query performance. Therefore, some other measures must be used in order to filter out non-useful terms and select suitable terms. Selecting suitable query terms is only one step toward improving query performance. In order to optimize query performance some parameters are to be set: number of terms to be added to query, number of top ranked documents used for selecting query terms. In absence of any theoretical justifications these parameters have to be set empirically.

In this paper we have suggested some measures to improve efficiency of co-occurrence based query expansion. We have suggested use of information theoretic approaches to rank the co-occurring terms. One of the approaches used is Kullback-Liebler Divergence (KLD) and other is the variant of KLD. Extensive experiments have been done to adjust the parameters (number of terms to be added to query, number of top ranked documents). The results have been compared and analyzed for all the three methods.

In the rest of this paper, we first make a review on related work in Section 2. 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].

Some work on query expansion has been based on probabilistic models of the retrieval process. Researchers have 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 Approach

The methods based on the term co-occurrence which have been used since the 70's to identify the semantic 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 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."

The main problem with the co-occurrence approach was mentioned by 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. 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. We have used the pseudo relevance feedback method where we select top N documents using cosine similarity measures and terms are selected from this set.

In order to select co-occurring terms we have used two well-know coefficients: - jaccard and frequency, which are as follows.

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

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.

$$freq\_co(t_i, t_j) = \sum_{d \in D} (f_{d,t_i} \times f_{d,t_j}) \tag{2}$$

t_i = *all terms of top N docs terms*
t_j = *query terms*
$f_{d,ti}$ = *frequency of term* t_i *in doc*
$f_{d,tj}$ = *frequency of term* t_j *in doc*
d = *top N doc*

We apply these coefficients to measure the similarity between terms represented by the vectors. However, there is a risk in applying these measures directly, since the candidate term could co-occur with the original query terms in the top documents 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. The larger the number of co-occurrences, the less likely that term co-occur with query terms by chance. In order to reduce probability of adding the term by chance, we use the following equation to measure the degree of co-occurrence of a candidate term with query.

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

Where

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

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(t_i,t_j) or freq_co(t_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 $t_1, t_2 ... t_n$. For this suitabilityforQ is computed.

$$SuitabiltyforQ = f(c, Q) = \prod_{t_i in Q} (\delta + co\_degree(c, t_i))^{idf(t_i)} \tag{5}$$

To expand a query Q, we rank the terms in the top ranked documents according to their suitability for Q and choose the top ranked terms for query expansion.

In general the co-occurrence based approach selects highly frequent co-occurring terms with respect to the query terms. However, good query expansion terms are those terms that are closely related to the original query and 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. Information theoretic approaches have been found useful to incorporate above-mentioned idea. Next section deals with use of information theoretic approach for query expansion.

4 Information-Theoretic Approach

Information theoretic approaches used in query expansion are 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. 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:

$$KLD(t) = \left[p_R(t) - p_C(t) \right] \log \frac{\frac{f(t)}{NR}}{p_C(t)} \tag{6}$$

Here $P_R(t)$ is 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, analogously to $P_R(t)$;

$$P_R(t) = \begin{cases} \gamma \frac{f(t)}{NR} if t \in V(R) \\ \delta p_C(t) otherwise \end{cases} \tag{7}$$

Where
c is the set of all documents in the collection
R is the set of top retrieved documents relative to a query.
V(R) is the vocabulary of all the terms in R.
NR is the number of terms in R.
f(t) is the frequency of t in R

We have done our experiments with one more variation in which we have used a function other than f(t)/NR, taking also into account the likely degree of relevance of the documents retrieved in the initial run:

$$KLD\_\text{variation}(t) = \left[p_R(t) - p_c(t) \right] \log \frac{\dfrac{\sum_d f(t) \times score_d}{\sum_t \sum_d f(t) \times score_d}}{p_c(t)} \tag{8}$$

In order to see the effect of information theoretic measures, we first selected the expansion terms using suitability value (equation 5) then equation (6 and 8) was used to rank the selected terms. For calculating the value of $P_R(t)$(equation 7) we set γ=1, which restricts the candidate set to the terms contained in R. and then the top ranked terms for query expansion.

5 Description of Our Methodology

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

1. *Indexing* - Our system first identified the individual terms occurring in the document collection.
2. *Word stemming*. To extract word-stem forms, we used porter-stemming algorithm [20].
3. *Stop wording*. We used a stop list to delete the common occurring words from the documents.
4. *Document weighting*. We assigned weights to the terms in each document by the classical *tf.idf* scheme.
5. *Weighting of unexpanded query:* To weigh terms in unexpanded query, we used the *tf* scheme.
6. *Document ranking with unexpanded query*: We computed a document ranking using common coefficients jaccard between the document vectors and the unexpanded query vector.
7. *Listing of candidate terms:* We use jacc_coefficient or freq_coefficient using equation (1) or (2) to list out the candidate terms which could be used for expansion.
8. *Expansion term ranking*: The candidates were ranked by using equation (5) or (6) and top terms were chosen for expansion.
9. *Construction of expanded query*: We simply added the top terms to the original query.
10. *Document ranking with expanded query*: The final document ranking was computed by using jaccard coefficient between the document vectors and the expanded query vector.

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 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, equation (5) was used for selecting the expansion terms in ranked order. In the second approach, we selected all the terms based on suitability (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 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, and R-Precision. 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 documents from the top ranked set 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.

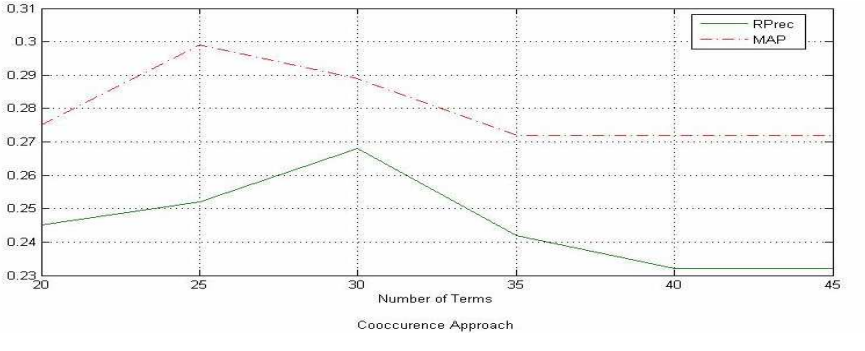

Fig. 1. Curve showing the MAP and R-PREC measures with different numbers of candidate terms to expand the original query using Co-occurrence Approach

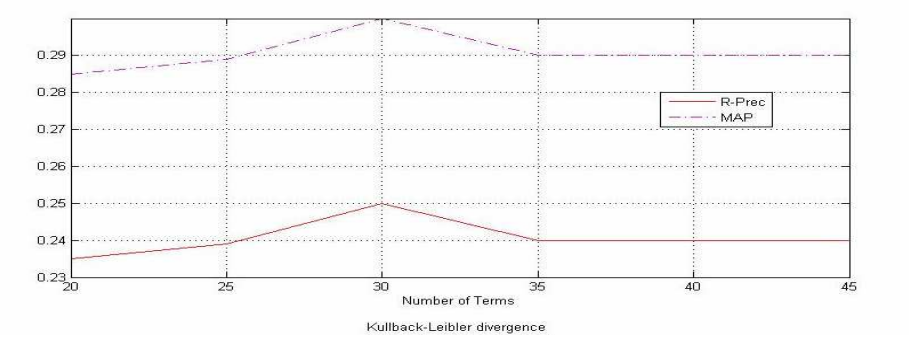

Fig. 2. Curve showing the MAP and R-PREC measures with different numbers of candidate terms to expand the original query using Kullback-Leibler divergence Approach

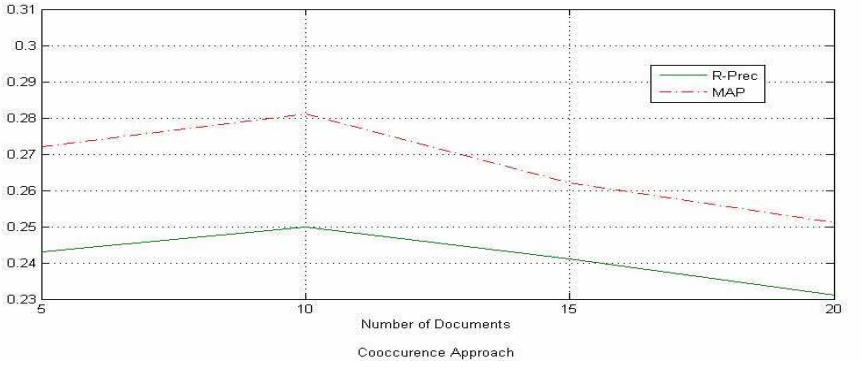

Fig. 3. Curve showing the MAP and R-PREC measures with different numbers of top documents used to extract the set of candidate query terms

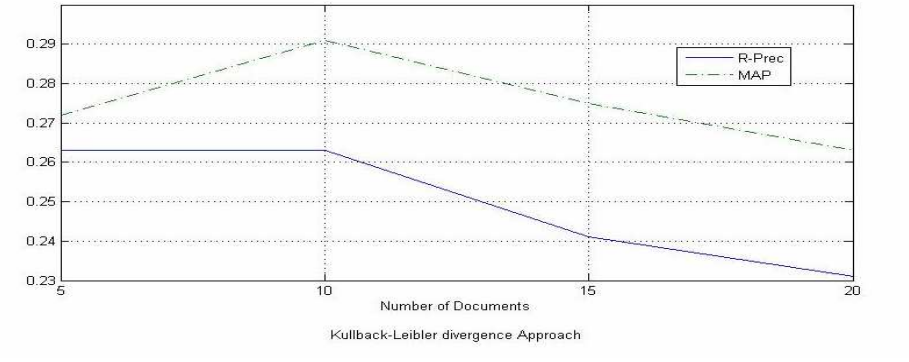

Fig. 4. Curve showing the MAP and R-PREC measures with different numbers of top documents used to extract the set of candidate query terms

We can observe that in all cases the best value for number of document selected for query expansion is around 10 documents and for the number of query expansion terms is 30. 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 10 and number of query terms to be added are 30. From the table we can observe that in general terms selected with suitability ranking are better candidates for query expansion in comparison to standard jaccard and frequency coefficients. We also observed that with the KLD we are able to improve the overall precision (MAP) and recall. 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. Best results appear in boldface.

	MAP	P@5	P@10	R-Prec
Unexpanded query approach	.2413	.3220	.2915	.2422
Jaccard_coeffcient	.2816	.3450	.2900	.3102
Freq_coeffiecient	.2218	.3146	**.2995**	.3018
Candidate term ranking using Suitabiltyof Q	.2772	.3660	.2820	.3643
Candidate term ranking using KLD	**.3012**	.3640	.2860	**.3914**
KLD_variation	.2970	**.3665**	.2840	.2802

7 Conclusions and Future Works

In this paper we have suggested the use of information theoretic measures in order to improve efficiency of co-occurrence based automatic query expansion. The experiments were performed on TREC dataset. We have used standard KLD as one of the information theoretic measures and suggested a variant of KLD. We observe that there is a considerable scope of improving co-occurrence based query expansion by using information theoretic measures. More experiments can be done in order to visualize the effect of suggested KLD variant. Further, the other information theoretic measures can be proposed to improve efficiency of automatic 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. Van Rijsbergen, C.J.: A theoretical basis for the use of cooccurrence data in information retrieval. Journal of Documentation (33), 106–119 (1977)
3. Carpineto, C., Romano, G.: TREC-8 Automatic Ad-Hoc Experiments at Fondazione Ugo Bordoni,TREC (1999)
4. Croft, W.B., Harper, D.J.: Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35, 285–295 (1979)
5. 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)
6. 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)
7. Efthimiadis, E.N.: Query expansion. Annual Review of Information Systems and Technology 31, 121–187 (1996)
8. Voorhees, E.M.: Overview of the TREC 2003 robust retrieval track. In: TREC, pp. 69–77 (2003)
9. Voorhees, E.M.: The TREC 2005 robust track. SIGIR Forum 40(1), 41–48 (2006)
10. Voorhees, E.M.: The TREC robust retrieval track. SIGIR Forum 39(1), 11–20 (2005)
11. Cao, G., Nie, J.Y., Gao, J.F., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proc. of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250 (2008)
12. Imran, H., Sharan, A.: Thesaurus and Query Expansion. International journal of computer science & information Technology (IJCSIT) 1(2), 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, 189–216 (1978)
14. Peat, H.J., Willett, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems. JASIS 42(5), 378–383 (1991)
15. Schütze, H., Pedersen, J.O.: A cooccurrence-based thesaurus and two applications to information retrieval. Inf. Process. Manage 33(3), 307–318 (1997)
16. Jing, Y., Croft, W.B.: An association thesaurus for information retrieval. In: 4th International Conference on Proceedings of RIAO 1994, New York, US, pp. 146–160 (1994)
17. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. 18(1), 79–112 (2000)
18. Lesk, M.E.: Word-word associations in document retrieval systems. American Documentation 20, 27–38 (1969)
19. 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)
20. Porter, M.F.: An algorithm for suffix stripping. Program - automated library and information systems 14(3), 130–137 (1980)
21. Maron, M.E., Kuhns, J.K.: On relevance, probabilistic indexing and information retrieval. Journal of rhe ACM 7, 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, 329–348 (1972)

23. 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)
24. 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)
25. Mandala, R., Tokunaga, T., Tanaka, H.: Ad hoc retrieval experiments using wornet and automatically constructed theasuri. 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, 129–146 (1976)
27. Liu, S., Liu, F., Yu, C., Meng, W.: 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, 239–246 (1983)
30. Sparck Jones, K.: Automatic keyword classification for information retrieval. Butterworth, London (1971)
31. Van Rijsbergen, C.J., Harper, D.J., Porter, M.F.: The selection of good search terms. Information Processing and Management 17, 77–91 (1981)
32. Qiu, Y., Frei, H.-P.: Concept based query expansion. In: SIGIR, pp. 160–169 (1993)