Comparison of Global Term Expansion Methods for Text Retrieval

Yuen-Hsien Tseng, Yu-Chin Tsai*, and Chi-Jen Lin**
National Taiwan Normal University, Taipei, Taiwan, R.O.C., 106
samtseng@ntnu.edu.tw
*Fu Jen Catholic University, Taipei, Taiwan, R.O.C., 242
tlinux@mail.my-net.idv.tw
**WebGenie Information LTD., Taipei, Taiwan, R.O.C., 106
dan@webgenie.com.tw

Abstract

This paper describes our work at the fifth NTCIR workshop on the subtasks of single language information retrieval (SLIR). Several automatic global query expansion strategies were explored based on a machine-derive thesaurus. These term selection strategies were compared with manual selection and local expansion. Experiments show that all the global expansion strategies perform worse than the simple local expansion. Furthermore, even with the help of a human in selecting the global terms, the performance may not be better than an automatic local feedback method, if the human does not fully understand the information need of the search topic. However, the machine-derived thesaurus does extract relevant global terms for expansion. Proper term selection can yield great improvement in performance than local feedback alone.

Keywords:

Chinese IR, term association, global expansion.

1. Introduction

In NTCIR-3, we participated in the Chinese, Japanese, and Korean single-language retrieval tasks (SLIRs) using an information retrieval system that dealt with these three languages in exactly the same way without using language-dependent knowledge or resources [1]. Results showed that our retrieval effectiveness does not show any difference among these three SLIRs. However, the effectiveness of our system was lower than the average of all runs submitted to the NTCIR by all participants. Post-analysis showed that the basic retrieval strategies used in our and others' systems did not match top-performing systems that used other sophisticated techniques such as blind relevance feedback (BRF), probabilistic retrieval model, hybrid term indexing, and title words re-weighting [2-4]. Among these sophisticated techniques, BRF is the major approach that improves performance most.

Relevance feedback is a technique that modifies the original query based on the initial retrieval results. If relevant (or irrelevant) terms can be identified from the initial results, adding them to (or subtracting them from) the original query for another run of retrieval often improves the retrieval effectiveness. However, since relevant terms are unknown to the system until the initial results were inspected and feedbacked by a searcher, most BRF methods under automatic retrieval mode simply assume that the top-ranked documents retrieved from the initial query are relevant. Terms are then extracted from these "relevant" documents so as to add to the initial query.

This way of query modification or query expansion is called "local expansion" since only a handful of documents "relevant" to the initial query are used. Information from the rest of the documents is not exploited at all. In contrast, if the feedback information comes from the entire collection, we call this way of relevance feedback "global expansion". Our preliminary experiments in NTCIR-4 showed that if best term selection can be achieved, query expansion based on the feedback terms from the entire collection can perform similarly well with local BRF and combining both local and global expansion can outperform each method alone [5].

A number of global expansion methods has been proposed in the past literature. A common way is to use a thesaurus, which is a powerful search aide tool heavily referred to in information access services. A thesaurus lists candidate search terms and relationships among them. This vocabulary knowledge can be used for broadening or narrowing the user's search topics, or for suggesting synonyms or related terms to reduce the vocabulary mismatch problem, which has long been one of the major

causes of search failure. However, tools like this often require laborious human involvement to add knowledge to them, making using thesauri an expensive choice. Besides, the use of general purpose thesauri does not improve retrieval effectiveness in a number of information retrieval experiments [6-8]. Query expansion based on manual thesauri only succeeds when the terms covered by the thesauri correspond closely to the vocabulary used in the document collection [9-10]. Thus, thesauri that are collection-specific and that can be easily generated and maintained with minimum cost are of great value.

For term suggestion and collection exploration during an interactive search scenario, we have proposed an automatic and efficient way for generating a thesaurus based on term co-occurrence. We believe that if best thesaurus terms can be selected, the thesaurus can also help in automatic search scenario such as in those topic-based retrieval tasks of NTCIR workshops.

This work reports our efforts in selecting the thesaurus terms for global expansion. We experimented on the collections of NTCIR-3 and applied the best results to the collections of NTCIR-5. Next section will introduce our ways of selecting global terms for query expansion. Section 3 will describe the retrieval strategies and experimental results on NTCIR-3's collections. Section 4 reports our retrieval results submitted to the NTCIR-5. Finally we conclude and summarize our observations in Section 5.

2. Selection of Global Relevant Terms

There are a number of approaches to generate a co-occurrence thesaurus from the entire document collection. However, most methods require a tremendous amount of computation. Specifically, $O(m^2)$ term-to-term similarities were calculated with each term-pair similarity requiring $O(n)$ steps, where m is the number of collection terms and n is the number of collection documents. This leads to an $O(m^2n)$ method. Due to this difficulty in obtaining a thesaurus, Tseng proposed another method that is far more efficient [11]. The major idea of his method is to limit the terms to be associated to those that co-occur in the same logical segments of a smaller text size, such as a sentence or a paragraph, rather than in the entire document. Association weights are computed in this way for each document and then accumulated over all documents. This changes it into a roughly $O(nk^2s)$ algorithm, where k is the number of selected keywords for association in a document and s is the average number of sentences in a document. In this way, a global term relation structure can be obtained efficiently. For the 381,375 Chinese documents in the NTCIR-4 collection (469 MB of texts), it only takes 133 minutes on a notebook computer with a 1.7 GHz CPU and 512 Mega RAM for indexing, keyword extraction, and term association computation.

As to the effectiveness, two experiments have been conducted. In one experiment, 30 topics (single query terms) were selected from the index terms of 25,230 Chinese news articles, from which term association was analyzed. Five assessors (all majored in library science) were invited for relevance judgment. For each topic, its top N (N = 50) related terms were examined. Users were asked if they thought the relationship between a topic and each of its associated terms was relevant enough. If they are not sure (mostly due to lack of domain knowledge), they are advised in advance to retrieve those documents that may explain their associations. The results show that in terms of percentage of relatedness, there are 69% associated terms judged relevant to the topic terms [11]. In another similar experiment based on a much larger collection (154,720 documents), the percentage of relatedness increases to 78.33% [12]. As can be seen, the more the documents for analysis, the better the effectiveness of the term association.

To have an idea of the term association results from the NTCIR collections, Figure 1 shows an example. The highlighted search term: "Akira Kurosawa" is from the topic 012 of NTCIR-4's CLIR topic set. Twelve top-ranked co-occurred terms were shown, with more 'relevant' terms in closer distance to the search term. In our implementations, at most 64 co-occurred terms for each keyword were kept in the term relation structure for later use.

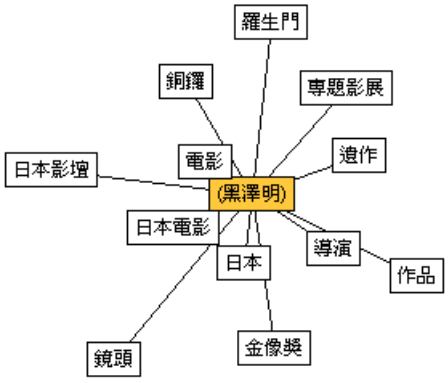

Figure 1. An example of global term expansion.

The example from Figure 1 seems to suggest that the associated terms are all good for expansion. But this is only a positive example. The associated terms indeed are related to the search terms with

respect to some topics or events mentioned in the collection. But some of such topics or events may not relate to the search topic represented by the single search term. Therefore, the goal in this work is to add to the query with proper terms that are associated with a set of search terms which together represent a search topic.

Specifically, each query string was segmented by the top-2 longest index terms. The resultant query terms, called keywords (KWs) hereafter, are used to fetch their related terms (RTs) from the thesaurus generated above. Our goal is to select the proper terms from these RTs for global expansion.

Four automatic strategies for global term selection have been tried. They are denoted as from S1 to S4 and are described as follows.

S1: Candidate RTs are selected based on their document frequencies relative to their corresponding keywords and based on whether they are common RTs of more than one KW. Specifically, high-frequency KWs (whose df > 20/n) were discarded. RTs having larger df than their KWs are also discarded. Finally, only those RTs associated to at least two KWs are used for expansion. This limitation on the added terms is to avoid topic drift, a phenomenon that changes the topic of the original query as more (irrelevant) terms were added. But this also limits the number of terms for global query expansion such that most topics have only a few additional terms. Actually this is the strategy that was used in our experiment in NTCIR-4.

S2: This strategy uses only common RTs without the frequency limitation. That is, common RTs of original KWs are used to retrieve new RTs. The intersection of the new RTs and the original RTs are used for expansion. Expanding RTs additionally may expand the topic foci. But by selecting the common RTs, we may have the chance to limit the focus of the topic to those that we need while broadening the base for RT candidates. Indeed, we have observed more additional RTs included by this strategy than those by S1.

S3: The above two strategies do not use the association strength information between the KWs and RTs. To calculate the strength of an RT with respect to the search topic, the asymmetric dice coefficient is used:

$$S(RT) = \sum_{KW \in Query} \frac{df(RT \cap KW)}{df(RT)}$$

The RTs are ranked by this strength in descending order. The best k RTs are then chosen for expansion. By varying k from 5, 10, 20, 30, to 40, our experiments showed that the best performance was achieved when k=5.

S4: In another attempt to know whether the RTs are relevant to the search topic, we make another document search by each RT and check whether titles of the top-ranked k documents contain the original KWs. If yes, the RT is included for expansion. Again, by varying k from 5, 10, 15, to 20, the best performance was observed when k=15.

Finally, to have an idea of how these RTs help in NTCIR's topic-based retrieval task, human judgment were conducted in comparison with the above automatic methods. From the Chinese SLIR in NTCIR-3, over 200 keywords were extracted from the descriptions of the 42 topics. Based on these keywords, a total of 8144 RTs were extracted from the thesaurus. Each of these 8144 RTs was then judged manually relevant or not to the search topics based on the topic description field. However, our preliminary experiment showed that the manual selected RTs from this judgment did not improve performance significant as expected. A second run of judgment was conducted. This time the RTs have to appear in the relevant documents of the search topics so as to be judged relevant or not. By limiting the RTs in this way, the judgment of these RTs can be more topic-specific and thus accurate.

3. Results for the NTCIR-3 Collection

Based on the Chinese SLIR collection of NTCIR-3, a number of experiments were conducted with various strategies (with or without global and local expansion) under different retrieval models. The used retrieval models are BS (byte size normalization), pivoted normalization method, BM11, BM25, and BM25m (modified BM25).

The BS method is an approximation of the cosine method as follows [13]:

$$BS(d_i, q_j) = \frac{\sum_{k=1}^{T} d_{i,k} q_{j,k}}{(bytesize_{d_i})^{0.375} \sqrt{\sum_{k=1}^{T} q_{j,k}^2}}$$

where the *bytesize* denotes the number of bytes of a document. The document term weight $d_{i,k}$ in the above is calculated by the term frequency, i.e., *log(1+tf)*. The query term weight $q_{j,k}$ is calculated by the term frequency and the inverse document frequency, i.e., *log(1+tf)* x *log(1+n/df)*, where n is the collection size. BS is the fast method among these 5 retrieval methods and was used in our NTCIR-3 and NTCIR-4 experiments.

The pivoted method is an improvement of the BS method proposed by Singhal et al [14]:

$$Pivot(d_i, q_j) = \sum_{k=1}^{T} tf_{j,k} \log\left(\frac{n+1}{df_k}\right) \left(\frac{1+\log(1+\log(tf_{i,k}))}{(1-s)+s\frac{dl_i}{Avgdl}}\right)$$

where T is the number of query terms, dl_i is the document length of document i, *Avgdl* is the average

document length in the collection, and *s* is a parameter, which is 0.2 in our experiment.

The BM11 and BM25 are probabilistic retrieval models that compute the okapi weight of a document with respect to a query:

$$BM11(d_i, q_j) = \sum_{k=1}^{T} tf_{j,k} \log\left(\frac{n - df_k + 0.5}{df_k + 0.5}\right)\left(\frac{tf_{i,k}}{tf_{i,k} + \frac{dl_i}{Avgdl}}\right)$$

$$BM25(d_i, q_j) = \sum_{k=1}^{T} \frac{(k_3+1)tf_{j,k}}{k_3 + tf_{j,k}} \log\left(\frac{n - df_k + 0.5}{df_k + 0.5}\right)\left(\frac{(k_1+1)tf_{i,k}}{tf_{i,k} + k_1((1-b) + b\frac{dl_i}{Avgdl})}\right)$$

where k_1, k_3 and b are parameters and are set as k_1=1.2, k_3=1000, and b=0.75 in our experiment. BM25 often outperforms BM11 for English documents. However, the (*n-df+0.5*) term may lead to unexpected effect for high-frequency terms, as was analyzed by Fang et al [15]. Therefore, they proposed a modified BM25 method by eliminating this effect as follows:

$$BM25m(d_i, q_j) = \sum_{k=1}^{T} \frac{(k_3+1)tf_{j,k}}{k_3 + tf_{j,k}} \log\left(\frac{n + 0.5}{df_k + 0.5}\right)\left(\frac{(k_1+1)tf_{i,k}}{tf_{i,k} + k_1((1-b) + b\frac{dl_i}{Avgdl})}\right)$$

As to the local expansion, thirty best terms from six top-ranked documents retrieved by the initial query were used. These six documents were first concatenated into one text string and then the keyword extraction algorithm [11] was applied to extract maximally repeated patterns. The extracted terms were sorted in decreasing order of occurrence. The first 30 terms were then selected for local query expansion. The decision on the number of best terms and the number of top-ranked documents was quite arbitrarily. We chose these numbers from the beginning almost without any tuning.

Table 1. Performance of various expansion strategies under different retrieval models for the Chinese SLIR description run.

	BS	Pivot	BM11	BM25	BM25m
Basic	0.2355	0.1921	0.2329	0.2422	0.2415
L	0.2876	0.2087	0.3028	0.3035	0.3080
S1	0.2512	0.1891	0.2482	0.2514	0.2502
S2	0.2505	0.1793	0.2482	0.2523	0.2509
S3	0.2524	0.2224	0.2511	0.2570	0.2571
S4	0.2134	0.1705	0.2136	0.2184	0.2183
H1	0.2929	0.2441	0.2877	0.2932	0.2930
H2	0.3299	0.2746	0.3286	0.3327	0.3329
S3+L	0.2960	0.2467	0.3092	0.3141	0.3133
H2+L	0.3431	0.2840	0.3562	0.3543	0.3544
Max of C-C-D in NTCIR-3				0.4990	
Avg of C-C-D in NTCIR-3				0.2670	

Table 1 shows our experiment results for the Chinese SLIR task in NTCIR-3. The query string is from the description field of the search topic. The 'Basic' row denotes a baseline result without any expansion. The 'L' row denotes the result from local expansion. S1 to S4 are those from global expansion strategies S1 to S4, respectively. H1 and H2 are those results obtained by adding the global terms judged relevant manually. H1 denotes our first judgment attempt. H2 denotes the second, as is explained in the above. As to the last two rows, they denote the combined strategies by using the global term expansion first, followed by the local expansion.

As Table 1 shows, all these four strategies perform worse than the simple local expansion. However, if proper terms can be selected from the suggested RTs, the performance can be better than the local feedback, as shown by the results of H2. Furthermore, combining both local and global expansion can outperform each method alone. This shows that the automatically generated thesaurus did extract relevant global terms for expansion. It is our automatic methods that fail to select good enough terms, despite we have tried out four different strategies.

Also note that the results of H1 are better than the local expansion for the BS and Pivot methods; but they are slightly worse for all the probabilistic models. This implies that even with the help of a human in selecting the expansion terms in an interactive search mode, the performance may not be better than a fully automatic feedback method, if it is equipped with a high-performing retrieval model. Only when the human fully understands what he/she needs (like the judgment in H2), can he/she outperform a machine-driven feedback method.

Another observation from Table 1 is that BM25 did perform better than the others. However, the pivoted normalization method did not perform well as expected. This may due to the improper parameter setting for this collection. But it also reveals its sensitivity in parameter tuning for obtaining a good result.

4. Results for the NTCIR-5 Collections

In NTCIR-5, the document collections are new and larger than those in NTCIR-3 and -4. The query topics prepared by NTCIR-5 consist of *title*, *description*, *narrative*, and *concept* fields. A total of about 50 topics for SLIR were provided for retrieval. Mean average precision (MAP) for these topics was calculated by the well-known trec_eval algorithm. Participants can submit multiple results, each comes from the run that uses different fields of query topics and/or different retrieval strategies to see how MAP

changes. Each submitted run is evaluated in two criteria, one is *relax*, meaning that the relevance judgment is done in a less strict way; the other is *rigid*, meaning that the relevance is judged in a more rigorous sense.

Our results for the SLIR (Chinese, Japanese, and Korean) were shown in Table 2. In the RunID, the letter T denotes the run that submits the *titles* as queries, D denotes the *descriptions*, and C the *concepts*. The used retrieval models are shown in the parentheses. All the runs apply the global expansion strategy S3 followed by the local expansion method.

Table 2. Performance of different runs.

RunID	Rigid	Relax
C-C-T (BS)	0.2599	0.3146
C-C-T (BM11)	0.2604	0.3217
C-C-D (BS)	0.2706	0.3270
C-C-D (BM11)	0.2820	0.3425
C-C-C (BM11)	0.2879	0.3622
Max of C-C-T	0.5047	0.5441
Avg of C-C-T	0.2874	0.3319
Max of C-C-D	0.4826	0.5249
Avg of C-C-D	0.2986	0.3523
J-J-T (BS)	0.1654	0.2333
J-J-T (BM11)	0.1514	0.2249
J-J-D (BS)	0.1825	0.2543
Max of J-J-T	0.4193	0.5028
Avg of J-J-T	0.2954	0.3795
Max of J-J-D	0.3823	0.4707
Avg of J-J-D	0.2861	0.3741
K-K-T (BS)	0.3044	0.3492
K-K-T (BM11)	0.2878	0.3250
K-K-D (BS)	0.3106	0.3531
K-K-D (BM11)	0.2750	0.3232
K-K-C (BM11)	0.3106	0.3357
Max of K-K-T	0.4912	0.5441
Avg of K-K-T	0.3846	0.4259
Max of K-K-D	0.5079	0.5680
Avg of K-K-D	0.4017	0.4507

As can be seen from Table 2, the probabilistic retrieval model performs slightly better than the vector space model regardless of long or short queries for the Chinese SLIR. But it performs worse for the Japanese and Korean SLIR. The reason may be due to the poor indexing schemes for the Japanese and Korean documents, for which our term indexing method do no use any vocabulary knowledge from these two languages, as was explained in our previous report [1].

5. Conclusions

We have tried several global expansion strategies based on an automatically generated thesaurus. Although the thesaurus does capture the relevant global terms for expansion, as shown in our human selection experiment, our automatic selection strategies do not come out with desired results. The difficulties lie in that among the 8144 related terms in the NTCIR-3 Chinese topics, only 214 are highly relevant to these topics, a ratio of only 2.63% (4252 are vaguely relevant and the remaining 3678 are irrelevant). The strategy S3 has a recall of 70/214=0.3271 and a precision of 70/(70+104)= 0.4223, while H2 has a recall of 162/214=0.7570 and a precision of 162/(162+809)=0.1668. Detailed analysis of these strategies was discussed in [16]. Generally speaking, the contexts of the related terms are important in judging their relevance. But these contexts are not easily found for a specific search topic, especially when the information need is constrained by some limitations in the narrative field of a search topic.

Nevertheless, we plan to explore more expansion strategies in the future to better exploit the co-occurrence thesaurus generated by the machine.

From this year's results and our past experiences, we conclude that the factors that affect the retrieval effectiveness are the quality of query terms, the term indexing schemes, the use of query expansion, retrieval models, and term weighting methods. Future work will consider these factors together to obtain better retrieval performance.

Acknowledge

This work is supported in part by NSC under the grant number NSC 94-2213-E-003 -012 -.

References

[1] Da-Wei Juang and Yuen-Hsien Tseng, "Uniform Indexing and Retrieval Scheme for Chinese, Japanese, and Korean," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.137-141.

[2] K. L. Kwok, "NTCIR-3 Chinese, Cross Language Retrieval Experiments Using PIRCS," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.45-49.

[3] Masaki Murata, Qing Ma, and Hitoshi Isahara, "Applying Multiple Characteristics and Techniques to Obtain High Levels of Performance in

Information Retrieval," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.87-92.

[4] Robert W. P. Luk, K. F. Wong, and K. L. Kwok, "Different Retrieval Models and Hybrid Term Indexing," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.93-100.

[5] Yuen-Hsien Tseng, Da-Wei Juang and, Shiu-Han Chen "Global and Local Term Expansion for Text Retrieval," Proceedings of the Fourth NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, June 2-4, 2004, Tokyo, Japan.

[6] Chen, H.-H., Lin, C.-C., and Lin, W.-C., "Construction of a Chinese–English wordNet and its application to CLIR," In Proceedings of the fifth International Workshop on information retrieval with Asian languages, pp. 189–196, 2000.

[7] Smeaton, A.F., & Berrut, C., "Thresholding postings lists, query expansion by word–word distances and POS tagging of Spanish text. In Proceedings of the fourth text retrieval conferences., 1996

[8] Voorhees, E.M., "Query expansion using lexical-semantic relations," Proceedings of the 17th ACM SIGIR conference on research and development in information retrieval, pp. 61–69, 1994.

[9] Fox, E.A., "Lexical relations enhancing effectiveness of information retrieval systems," SIGIR Forum, 15(3), 6–36, 1980.

[10] Mandala, R., Tokunaga, T., & Tanaka, H., "Combining multiple evidence from different types of thesaurus for query expansion," In Proceedings of the 22nd ACM SIGIR conference on research and development in information retrieval, pp. 191–197, 1999.

[11] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[12] Jia-Yun Ye, *Evaluation of Term Suggestion in an Interactive Chinese Retrieval System*, Master Thesis, Department of Library and Information Science, Fu Jen Catholic University, 2004. (in Chinese)

[13] Amit Singhal, Gerard Salton, and Chris Buckley, "Length Normalization in Degraded Text Collections," Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 149-162.

[14] Amit Singhal, Chris Buckley, and Mandar Mitra, "Pivoted Document Length Normalization," Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland, August 1996, Pages: 21 – 29.

[15] Hui Fang, Tao Tao, and ChengXiang Zhai, "A Formal Study of Information Retrieval Heuristics," Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25 - 29 Sheffield, U.K., 2004, pp. 49-56.

[16] Yu-Chin Tsai, *Term-Selection for Query Expansion in Topic-based Information Retrieval*, Master Thesis, Department of Library and Information Science, Fu Jen Catholic University, 2005. (in Chinese).