Selecting Good Expansion Terms based on Google Similarity Distance

Jing Luo[1][2] Bo Meng[1] Xinhui Tu[2] Jinguang Gu[2]

[1] School of Computer, Wuhan University, China

[2]School of Computer Science and Technology,

Wuhan University of Science and Technology, China
luoluocat@ mail.wust.edu.cn
bmengwhu@sina.com
tuxinhui@mail.wust.edu.cn
simon@ mail.wust.edu.cn

Abstract—**In this paper, we propose a novel expansion terms selection model, in which Google similarity distance is adopted to estimate the relevance between query and candidate expansion terms. In previous method, expansion terms are usually selected by counting term co-occurrences in the documents. However, term co-occurrences are not always a good indicator for relevance, whereas some are background terms of the whole collection. In order to select good expansion terms, Google similarity distance is adopted in our model to estimate two kinds of relevance weight. One is the relevance weight between query and its relevant term extracted from the top-ranked documents in initial retrieval results. The other is the relevance weight between each query term and its relevant terms extracted from the snapshot of Google search result when that query term is used as search keyword. The estimated relevance weights are used to select good expansion terms for second retrieval. The experiments on the two test collections show that our expansion terms selection model is more effective than the standard Rocchio expansion.**

Keywords-Query expansion; relevant terms; information retrieval

I. INTRODUCTION

Usually, users describe their information needs by a few keywords in their queries, which are likely to be different from those terms in the documents. As a consequence, in many cases, the documents returned by information retrieval system are not relevant to the user information need. This raises a fundamental problem of term mismatch in information retrieval, which is also one of the key factors that affect the precision of the information retrieval system.

To solve this problem, various query expansion techniques were proposed in [1-5]. The basic idea of query expansion is to supplement the original query with additional terms which are relevant to the original query. There are two key aspects in any query expansion technique: the source where expansion terms are selected and the method to weight and integrate expansion terms.

Previous automatic query expansion techniques can be generally categorized into global analysis and local analysis.

The basic idea of global analysis is to use the context of a term to determine its similarity with other terms. Global analysis selects expansion terms from the whole document set. It builds a set of statistical term relationships which are then used to expand queries.

One of the earliest global analysis techniques is term clustering [6][7]. Queries are simply expanded by adding similar terms that are grouped into the same cluster according to term co-occurrences in documents. Qiu and Frei [8] presented a query expansion model using a global similarity thesaurus. Another work based on a global statistical thesaurus is [9], which first clusters documents and then selects low-frequency terms to represent each cluster.

Generally, global analysis requires corpus-wide statistics, such as statistics of co-occurrences of pairs of terms, resulting in a matrix of similarities between terms or a global association thesaurus. Although the global analysis techniques are relatively robust, the corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it focuses only on the document side and does not take into account the query side, global analysis only provides a partial solution to the term mismatching problem.

Different from global analysis, local analysis uses only a subset of returned documents with the given query. It is thus more focused on the given query than global analysis.

Local feedback has been proven effective in previous TREC experiments. In some cases, it outperforms global analysis [2][4][5][10][11]. Nevertheless, this method can hardly overcome its inherent drawback: If a large fraction of the top-ranked documents are actually irrelevant, then the words added into the query (drawn from these documents) are likely to be unrelated to the topic and as a result, the quality of the retrieval using the expanded query is degraded. Therefore, the effect of pseudo feedback strongly depends on the quality of the initial retrieval.

In previous method, expansion terms are usually selected by counting term co-occurrences in the documents. However, term co-occurrences are not always a good indicator for relevance, whereas some are background terms of the whole collection. In order to remove noise, we propose a novel expansion terms selection model in which Google similarity distance is adopted to estimate two kinds of relevance weight. One is the relevance weight between query and its relevant term extracted from the top-ranked documents in initial retrieval results. The other is the relevance weight between each query term and its relevant terms extracted from the

978-1-4244-5824-0/$26.00 ©2010 IEEE

snapshot of Google search result when that query term is used as search keyword. The estimated relevance weights are used to select good expansion terms for second retrieval.

The rest of this paper is organized as follows: Section 2 demonstrates the process of our IR system. Section 3 describes the term extraction algorithm used in our experiments. Section 4 describes the Google similarity distance measure used in our experiments. In section 5, we describe how to select good expansion terms in detail. The proposed query expansion procedure is described in Section 6. Section 7 evaluates the performance of this method on two NTCIR test collections and gives out some result analysis. Section 8 gives the conclusion and some future work.

II. System description

Fig.1 demonstrates the process of our Chinese information retrieval system. Following is the procedure of our retrieval system:

1) The short terms automatically extracted from document sets are used to build indexes, and the short terms in both the query and documents are used to do initial retrieval.
2) Two kinds of relevance weight are computed by Google similarity distance. The first one is the relevance between the query and the candidate expansion terms extracted from the top N documents in initial retrieval results. And the second is the relevance between each query term and the candidate expansion terms extracted from the snapshot of Google search result when each query term is used as search keyword.
3) The estimated relevance weights are used to select good topic-relevant terms for second retrieval.
4) The topic-relevant terms come from two kinds of source are used together to do query expansion.
5) The new query is used to do second retrieval and get final ranking documents.

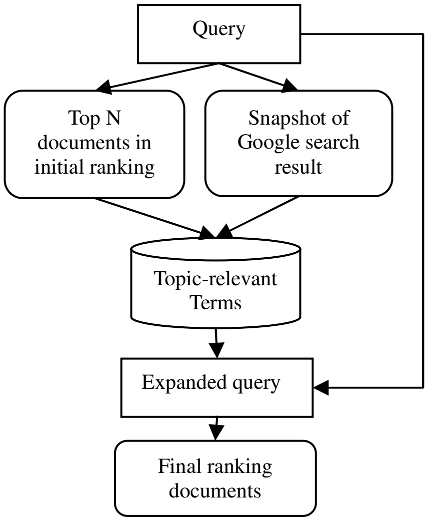
Query
Top N documents in initial ranking
Snapshot of Google search result
Topic-relevant Terms
Expanded query
Final ranking documents

Fig 1. Process of our system

III. Term extraction

For Chinese information retrieval task, bi-gram and word both are the most effective indexing units [12][13][14]. But they are not ideal units for query expansion. We use automatically extracted terms as expansion units in our work in order to improve effectiveness of query expansion.

Term extraction concerns the problem of what is a term. Intuitively, key terms in a document are some word strings which are conceptually prominent in the document and play main roles in discriminating the document from other documents.

A seeding-and-expansion mechanism proposed by [15] is adopted in our experiments to extract key terms from documents. We regard a term whose length is less than 4 Chinese Characters as a short term, and a term whose length is equal or greater than 4 Chinese Characters as a long term. The following are some examples of some short and long terms.

1) Short Terms:
基因 (gene), 食物(food), 蛋白质 (protein)
2) Long Terms:
老虎伍兹 (Tiger Woods), 奥林匹克 (Olympic), 欧洲货币组织 (European Monetary Fund)

IV. Google similarity distance

The world-wide-web is the largest database on earth, and the context information entered by millions of independent users averages out to provide automatic semantics of useful quality.

Google similarity distance [16] is a new theory of relevance between words and phrases based on information distance and Kolmogorov complexity. It use the world-wide-web as database, and Google as search engine. This theory is then applied to construct a method to automatically extract similarity, the Google similarity distance, of words and phrases from the world-wide web using Google page counts.

Suppose x and y are two different terms, N is the number of whole pages in Google web search engine, then the Google similarity distance between x and y can be computed by the following formula.

$$G(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}} \tag{1}$$

where *f* -*x*- -denotes the number of pages containing *x*, and *f* -*x*ψ*y*- -denotes the number of pages containing both *x* and *y*, as reported by Google.

While the theory is rather intricate, the resulting method is simple enough. We give an example: At the time of doing the experiment, a Google search for “基因”, returned 19,800,000 hits. The number of hits for the search term “食物” was 58,200,000. Searching for the pages where both “基因” and “食物” occur gave 2,420,000 hits, and Google indexed 25,200,000,000 web pages. Using these numbers in the main formula (1) we derive below, with N = 25,200,000,000, this yields a Normalized Google Distance between the terms “基因” and “食物” as follows:

G(“基因”, “食物”) = 0.444839

V. Expansion terms selection

A. Expansion terms from Google snapshots

To select the relevant terms from Web for each term in query, Google similarity distance is used to estimate the relevance weight between a query term and each term extracted from the snapshots returned by Google when we use the query term as search keyword.

For each query term q_k, we can extract its relevant terms from the snapshots returned by Google search engine, with the query term as search keyword. The term extraction procedure is the same as we described in section 4. The corresponding Google snapshot of query term q_k is define as document d_k.

Intuitively, we can use the occurrence of terms in the document d_k to determine the relevance weight between a query term and its relevant terms. If all terms appearing in the document d_k center on the topic of search term q_k, then we can simply use a maximum likelihood estimator and the problem is as simple as frequency counting. However, some terms in Google snapshots are not truly relevant to the search term, whereas some are background terms of the web collection.

In our experiments, Google similarity distance is used to estimate the relevance between query and candidate expansion terms. We regard term y is relevant to query term x if $G(x, y) \geqslant \lambda$ (λ >1). That is, the Google similarity distance between term x and y must higher than a threshold.

To select those terms that are most relevant to the concept of the query, rather than select terms that are only relevant to one of the query terms, we adopt a topic-relevant terms selection method similar to the one proposed in [3].In earlier methods, terms are selected that are strongly relevant to one of the query terms. The methods differ in the kind of relationships used. The entire query - in other words, the query concept - is seldom taken into account. This may be compared to translating from a natural language text into another: A dictionary look-up for a word does not give the final answer in many cases. Rather, the translator who knows the meaning of the text has to choose the suitable word from an entire list of possible translations. Likewise, we should consider a term that is relevant to the query concept rather than one that is only relevant to a single term in the query.

For each query, we use the following formula to select topic-relevant terms for them. Suppose {q_1,……, q_m} is the set of query terms, w is a term we extracted from the snapshots returned by Google.

$$R(w)=\frac{\sum_{i=1}^{m}G(w,q_i)}{|q|} \tag{2}$$

In practice, we only select the top N_g terms whose relevance weight to the query q is higher than other terms for query expansion.

B. Expansion terms from initial retrieval result

To select the relevant terms from initial retrieval result for the query, Google similarity distance is also used to estimate the relevance weight between the query and each term extracted from the initial retrieval result. For each query, we use the following formula to select topic-relevant terms for them. Suppose {q_1,……, q_m} is the set of query terms, w is a term we extracted from the top-ranked initial retrieval result.

$$R(w)=\frac{\sum_{i=1}^{m}G(w,q_i)}{|q|} \tag{3}$$

In practice, we only select the top N_d terms whose relevance weight to the query q is higher than other terms for query expansion.

VI. Query expansion

The topic-relevant terms extracted from two kinds of source and the terms in the query are used together to do query expansion. Following is the procedure to expand a query q:

1) For the terms from Google snapshot, we select top Ng topic-relevant terms and add them into q with R(w) as its weight.
2) For the terms from initial retrieval result, we select top Nd topic-relevant terms and add them into q with R(w) as its weight.
3) All terms in q are added to new query with frequency in q as its weight.

The original query plus new terms acquired by query expansion form a new query. This new query is used to search again to get final search result.

VII. EXPERIMENTS AND EVALUATION

In order to evaluate the effectiveness of the proposed method, we conducted experiments using the NTCIR-5 and NTCIR-6 CLIR Chinese sub collections. We use TITLE and DESCRIPTION fields in the query set. For keeping measurement granularity, each document is assigned one of four degrees of relevance id in the judgment process: S (highly relevant), A (relevant), B (partially relevant), or C(irrelevant). In the CLIR task, two different evaluation criterions are defined:

1) Rigid relevant: S+A
2) Relaxed relevant: S+A+B

In our experiments, the Okapi BM25 model [17][18] is used as baseline. For the BM25 model, the relevance between the document and the query is defined in the following formulas.

$$\sum_{t \in q} w_t \frac{(k_1+1)tf_d(t)}{K+tf_d(t)} \frac{(k_3+1)tf_q(t)}{k_3+tf_d(t)} \qquad (4)$$

$$w_t = \log \frac{(N-df(t)+0.5)}{df(t)+0.5} \qquad (5)$$

$$K = k_1 \times ((1-b)+b \times \frac{dl}{avdl}) \qquad (6)$$

where w_t is the Robertson/Spark Jones weight of t. k_1, b and k_3 are parameters. k_1 and b are set as 1.2 and 0.75 respectively by default, and k_3 is set as 7. dl and avdl are respectively the document length and average document length measured by the number of the index terms.

We also compare our method with query expansion using the traditional relevance feedback technique, in which the top 30 documents in initial retrieval result are used for feedback. We use the Rocchio formula [19] for term reweighing as follows:

$$Q_{new} = \alpha Q_{old} + \beta \sum_{r=1}^{n_{rel}} \frac{D_r}{n_{rel}} - \gamma \sum_{n=1}^{n_{norel}} \frac{D_n}{n_{norel}} \qquad (7)$$

where α, β and γ, are constants, Dr is the vector of a relevant document d_r, D_n is the vector of an irrelevant document d_n, n_{rel} is the number of relevant documents retrieved, and n_{nonrel} is the number of irrelevant documents. We set α=8, β=16, and γ=4 for this experiment.

In our experiments, the parameter M, N_g and N_d described in section 5 are set to 30, 30 and 30 respectively. This is, we select the 30 top ranked relevant terms from Google snapshot and the 30 top ranked relevant terms from the 30 top ranked documents in the initial retrieval to do query expansion.

The experiment results are given in Table 1 and Table 2. The tables show the mean average precision for each case, expansion using the relevance feedback technique and expansion using our method. For each method we give the percentage of improvement over the baseline. Experiment results show that the performance of our query expansion model is more effective than BM25 model and the standard Rocchio expansion.

VIII. CONCLUSIONS

In this paper we propose a novel term relevance estimation model to select good expansion terms for Chinese information retrieval. The experiments on two collections show that our query expansion model is more effective than the standard Rocchio expansion.

For the further work, we will try to improve the quality of topic-relevant terms and carry out experiments with this approach on other text collections to test its robustness.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 60803160).

TABLE 1. Comparison results on NTCIR-5 CLIR Chinese sub collection

Standard	Section	BM25	Rocchio expansion (RE)		Our method		
		MAP	MAP	% change over BM25	MAP	% change over BM25	% change over RE
Rigid	Title	0.2395	0.3262	+36.20%	0.3791	+36.82%	+16.22%
	Description	0.2137	0.3749	+75.43%	0.4120	+48.13%	+9.90%
Relax	Title	0.2843	0.3712	+30.57%	0.4325	+34.27%	+16.51%
	Description	0.2641	0.3946	+49.41%	0.4731	+44.18%	+19.89%

TABLE 2. Comparison results on NTCIR-6 CLIR Chinese sub collection

Standard	Section	BM25	Rocchio expansion(RE)		Our method		
		MAP	MAP	% change over BM25	MAP	% change over BM25	% change over RE
Rigid	Title	0.1537	0.2104	+36.89%	0.2437	+36.93%	+15.83%
	Description	0.1812	0.2310	+27.48%	0.2703	+32.96%	+17.01%
Relax	Title	0.2463	0.2978	+20.91%	0.3542	+30.46%	+18.94%
	Description	0.2528	0.3124	+23.58%	0.3696	+31.60%	+18.31%

REFERENCES

[1] G. Salton and C. Buckley, “Improving Retrieval Performance by Relevance Feedback,” J. Am. Soc. for Information Science, vol. 41, no. 4, pp. 288-297, 1990.

[2] C. Buckley, G. Salton, J. Allan, and A. Singhal, “Automatic Query Expansion Using SMART,” Overview of the Third Retrieval Conf. (TREC-3), pp. 69-80, Nov. 1994.

[3] Y. Qiu and H. Frei, “Concept Based Query Expansion,” Proc. 16th Int’l ACM SIGIR Conf. R & D in Information Retrieval, pp. 160-169,1993.

[4] J. Xu and W.B. Croft, “Query Expansion Using Local and Global Document Analysis,” Proc. 19th Int’l Conf. Research and Development in Information Retrieval, pp. 4-11, 1996.

[5] J. Xu and W.B. Croft, “Improving the Effectiveness of Information Retrieval with Local Context Analysis,” ACM Trans. Information Systems, vol. 18, no. 1, pp. 79-112, Jan. 2000.

[6] M.E. Lesk, “Word-Word Associations In Document Retrieval Systems,” Am. Documentation, vol. 20, no. 1, pp. 27-38, 1969

[7] K. Sparck Jones, “Automatic Keyword Classification for Information Retrieval”. London: Butterworths, 1971.

[8] Y. Qiu and H. Frei, “Concept Based Query Expansion,” Proc. 16th Int’l ACM SIGIR Conf. R & D in nformation Retrieval, pp. 160-169,1993.

[9] C.J. Crouch and B. Yang, “Experiments in Automatic Statistical Thesaurus Construction,” Proc. CM-SIGIR Conf. Research and Development in Information Retrieval, pp. 77-88, 1992.

[10] E. Efthimiadis and P. Biron, “UCLA-Okapi at TREC-2: Query Expansion Experiments,” Proc. Second Text Retrieval Conf. (TREC-2), D.K. Harmon, ed., 1994.

[11] S.E. Robertson, S. Walker, and M. Sparck Jones, et al., “Okapi at TREC-3,” Proc. Second Text Retrieval Conf. (TREC-3), 1995.

[12] Li, P.: “Research on Improvement of Single Chinese Character Indexing Method”. In: Journal of the China Society for Scientific and Technical Information, Vol. 18 No.5 (1999)

[13] Kwok, K.L., “Comparing Representation in Chinese Information Retrieval”, Proceeding of the ACM SIGIR-97, pp.34-41, 1997.

[14] Jian-Yun Nie, Fuji Ren, “Chinese information retrieval: using characters or words?”, Information Processing and Management, 35, pp.443-462 ,1999.

[15] Lingpeng Yang, DongHong Ji, Li Tang, “Document Re-ranking Based on Global and Local Terms”, Third SIGHAN Workshop on Chinese Language Processing, pp. 17-23 , 2004.

[16] RL.CILIBRAS, “The Google similarity distance”. IEEE Transactions on Knowledge and Data Engineering, Vol.19 No.3, pp.370-383, 2007.

[17] S.E. Robertson, S. Walker, and M. Sparck Jones, “Okapi at TREC-3”. Proc. of Third Text Retrieval Conference (TREC-3), 1995.

[18] Robertson, SE and Walker, S. and Beaulieu, M. “Experimentation as a way of life: Okapi at TREC”, Information Processing and Management, vol. 36, no.1 pp.95-108, 2000.

[19] Buckley C, Salton G, “The effect of adding relevance information in a relevance feedback environment”, Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 292-300, 1994.