Chinese Information Retrieval Based on Document Expansion

Tingting HE[1 2] Li LI[1 2] Guozhong QU[1 2] Yong ZHANG[1 2]
[1]Department of Computer Science, Huazhong Normal University, 430079, Wuhan
[2]Engineering Research Center of Education Information Technology Ministry of Education, 430079, Wuhan, China
hett@mail.ccnu.edu.cn heartlamp@gmail.com qu_g_z@mails.ccnu.edu.cn
ychang@mails.ccnu.edu.cn

Abstract

This paper describes our work at the sixth NTCIR workshop on the subtasks of monolingual information retrieval (CLIR). This is the second time we have participated in NTCIR. We have used query expansion methods in NTCIR-5 with related term groups, and this time we use document expansion. The traditional information retrieval model has limitations on finding related documents since it simply checks the existence of query terms in documents without considering the context of documents. Now we retrieve documents by vector space model and cluster the top-n documents to re-ranking the result set. Experiments show that our method achieves an average 3.2% improvement comparing with the method we have used in NTCIR-5 that adopts query expansion.

Keywords: *document expansion,cluster,information retrieval.*

1 Introduction

A lot of research has been done to improve retrieval effectiveness by using additional information about query or documents [1]. Two traditional methods are query expansion and document expansion. We use query expansion to improve the performance of Chinese information retrieval systems with related term groups in NTCIR-5 [2]. The new method fulfills document expansion, and achieves an average 10.7% improvement comparing with the traditional relevance feedback technique.

Firstly, we retrieve documents by traditional model. Second, we cluster the result set by group-average agglomerative method. Third, we calculate the similarity between document clusters and original query, and combine them as final results.

The paper is organized as following. In section 2, we describe the process of document expansion. In section 3, we evaluate the performance of our method and analyze the result. In section 4, we present the conclusion and some future work.

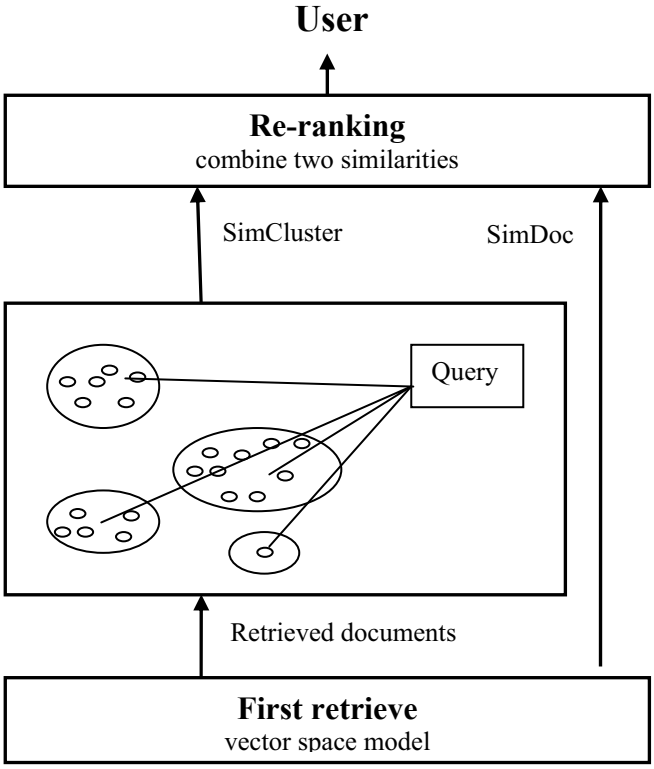

Figure 1. System structure

2 Processing

2.1 Result Set Cluster

We retrieve documents on the basis of the vector space model firstly. Then we get the top-n

documents as a result set to be clustered [3] (the value of n decided by the experiment results).

Most conventional clustering methods fall into two classes: non-hierarchical and hierarchical. Non-hierarchical clustering methods were first used because of their low computational requirements. These methods generally require a fixed number of clusters, which restriction makes them inappropriate for improving retrieval effectiveness. Recent applications of partitioning algorithms for information retrieval have also focused on issues of efficiency, rather than effectiveness. Hierarchical clustering methods have attracted much attention because they give the user the maximum amount of flexibility. Rather than requiring parameter choices to be pre-determined, the results represent all possible levels of granularity. So, most of the research on cluster analysis in information retrieval has employed hierarchical method [4]. We adopt the method of group-average agglomerative clustering, a compromise between single-link clustering and complete-link clustering.

We represent the objects as length-normalized vectors in an m-dimensional real-valued space, and define the similarity measure as cosine.

$$sim(\vec{x},\vec{y}) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{m} x_i \times y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} \times \sqrt{\sum_{i=1}^{m} y_i^2}} \quad (1)$$

For a cluster c_j, the average similarity S between vectors c_j is defined as follows: (The factor $|c_j|(|c_j|-1)$ calculates the number of non-zero similarities added up in the double summation.)

$$s(c_j) = \frac{1}{|c_j|(|c_j|-1)} \sum_{\vec{x}\in c_j} \sum_{\vec{x}\neq\vec{y}\in c_j} sim(\vec{x},\vec{y}) \quad (2)$$

If $s(\bullet)$ is known for two groups c_i and c_j, then the average similarity of their union can be calculated as follows:

$$s(c_i \cup c_j) = \frac{(\vec{s}(c_i)+\vec{s}(c_j))^2 - (|c_i|+|c_j|)}{(|c_i|+|c_j|)(|c_i|+|c_j|-1)} \quad (3)$$

If the similarity between the two clusters is greater than a certain threshold η, we merge the two clusters. η is the second factor determined by the experiment results.

2.2 Re-ranking

We calculate query-document similarity after the first retrieval and query-cluster similarity in the cluster analysis. That is, we focus on each document at the first step and on document collections at the second step [3].

We combine two similarities from the first retrieval and the second analysis step.

$$sim_{final} = sim_{doc} + \theta sim_{cluster} \quad (4)$$

Where θ is parameter to adjust the different values of the weighting schemes and give more importance to the similarities of the first or the second step. A document having low query-document similarity can be given high query-cluster similarity due to the effects of other documents in the cluster. In the reverse case, this is the same [4]. At the re-ranking step, we get the view matching the query by applying dynamic cluster partitioning to document of which similarity is calculated according to containment of query terms. And through the cluster analysis, the context of all terms in a document as well as query terms is considered. θ is the third factor we must get in our experiment.

3 Experiments and Evaluation

As a baseline, we used the SMART version 11.0 [5] system without query expansion. SMART is an information retrieval engine based on the vector space model in which term weights are calculated based on term frequency, inverse document frequency, and document length normalization [6]. The weighting method for document collection is as follows:

$$\frac{(\log(tf_{ik})+1.0)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij}+1.0)]^2}} \quad (5)$$

And the weighting method for the initial query is as follows:

$$\frac{(\log(tf_{ik})+1.0)*\log(N/n_k)}{\sqrt{\sum_{j=1}^{n}[\log(tf_{ij}+1.0)*\log(N/n_j)]^2}} \quad (6)$$

We compare our method with query expansion using the relevance feedback technique, in which 30 documents among the documents retrieved in the initial retrieval are used for feedback. We use the Rocchio formula for term reweighing as follows:

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

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.

First of all, we must confirm three factors by experiment: k, η, θ. k denotes the number of documents to be clustered, η is the threshold of the document clusters by group-average agglomerative method, and θ is the balance factor between the two similarities.

This is the experimental data，k is granted as 1000,1200,1400,1600,1800, respectively; η is granted as 0.5，0.6，0.7，0.8，0.9，respectively; θ is granted as 0.9，1.0，1.1，1.2，1.3，1.4，1.5，1.6，1.7, respectively.

Three optimal rigid description results of each group (the value of k is fixed) of experiments are listed below, with the answer of NTCIR-5 as criterion.

k	(η:θ)	precise	(η:θ)	precise	(η:θ)	precise
1000	(0.7:1.5)	35.021%	(0.7:1.4)	34.362%	(0.6:1.4)	34.227%
1200	(0.8:1)	38.617%	(0.7:1.1)	38.023%	(0.7:1)	37.689%
1400	(0.6:1.1)	40.721%	(0.6:1)	40.081%	(0.7:1)	38.674%
1600	(0.8:1.5)	38.172%	(0.8:1.4)	37.771%	(0.7:1.2)	37.294%
1800	(0.8:1.4)	36.982%	(0.8:1.3)	36.165%	(0.8:1.2)	35.667%

Table 1. Determine factors.

We get the following result: k=1400, η=0.6, θ=1.1. Now, compare the final result with that of NTCIR-5 in Table 2.

It shows the average precision of query expansion by related term groups method and our method. For each one, we give the percentage of improvement over the baseline method in parentheses. We can see that the performance of our method is better than SMART version 11.0 using the lnc.ltc term weighting method without expansion, and also better than expansion using the related term groups.

4 Conclusions

This paper proposes a novel method to improve the performance of Chinese information retrieval systems by expanding documents. Comparing to the method we used last year, the document clustering is added to traditional method instead of query expansion.

When we conducted experiments by NTCIR-5 method using the NTCIR-6 information retrieval test collections, we find that some irrelated terms are inserted into the query, and the effectiveness is more less than that is in NTCIR-5. Because NTCIR-6 needs the result set in a smaller domain, but the old method has to choose just one term group to expand query, that makes redundancy and document lost. We have considered the context of the documents partially and used document expansion in NTCIR-6. Experiments show that our method achieves an average 3.2% improvement.

References

[1] Jaroslaw Balinski, and Czeslaw Danilowicz, "Re-ranking method based on inter-document distances", Information Processing and Management 41(2005) 759-775

[2] Tingting HE, Guozhong QU, Xinhui TU, Donghong JI."Chinese information retrieval based on related term group". Proceedings of the Fifth NTCIR Workshop Meeting, page(s): 64-68, 2005

[3] Kyung-Soon Lee, Young-Chan Park, and Key-Sun Choi,"Re-ranking model based on document clusters", Information Processing and Management 37(2001) 1-14

[4] Gunhan Park, Yunju Baek, and Heung-kyu Lee, "Re-ranking algorithm using post-retrieval clustering for content-based image retrieval", Information Processing and Management 41(2005) 411－411

[5] Salton. G, *The SMART Retrieval System Experiments in Automatic Document Processing.* Englewood Cli.s,NJ: Prentice-Hall. 1971.

[6] Buckley C, and Salton G, "Automatic query expansion using SMART: TREC-3", *Overview of the Third Text Retrieval Conference (TREC-3),* (pp. 69-80). Gaithersburg, MD: NIST Special Publication 500-225, 1995.

[7] Yuk-Chi LI and Helen M. MENG, "Document Expansion using a Side Collection for Monolingual and Cross-language Spoken Document Retrieval", ISCA workshop on multilingual spoken document retrieval (MSDR2003) 7-7

[8] Xu J, and Croft B, "Query expansion using local and global document analysis". *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland (pp. 4-11). New York, NY: ACM Press 18-22 ,August 1996.

Standard	Base	Related term group		Our method		
	MAP	MAP	% Change over Base	MAP	% Change over Base	% Change over RTG
Rigid description	0.1664	0.2127	+27.8%	0.2354	+41.46%	+10.67%
Relax description	0.2419	0.3019	+24.8%	0.3294	+36.17%	+9.11%

Table 2. Comparison results on NTCIR-6 collection.

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

[10] Chen. H, Schatz. B, Yim. T, and Fye. D, "Automatic thesaurus generation for an electronic community system," *Journal of the American Society for Information Science*, 46(3), 175-193, 1995.

[11] Crouch. C. J, "An approach to the automatic construction of global thesauri," *Information Processing and Management*, 26(5), 629-640. 1990.

[12] Crouch. C, and Yang. B, "Experiments in automatic statistical thesaurus construction," *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark (pp. 77-82). New York, NY: ACM Press 21-24, 1992.

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

[14] Lee-Feng Chien.. "PAT-tree-based keyword extraction for Chinese information retrieval", *In Proceeding of the ACM SIGIR-97*. pp.50-58.,1997.

[15] Lee-Feng Chien.. "PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval.",*Information Processing and Management*, 35 (1999), 501-521, 1999.

[16] Buckley. C, and 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, July 1994.

[17] Buckley C, and Salton G, "Automatic query expansion using SMART: TREC-3," *Overview of the Third Text Retrieval Conference (TREC-3),* pp. 69-80, Gaithersburg, MD: NIST Special Publication 500-225, 1995