Re-ranking Method Based on Topic Word Pairs [1]

Tingting He[1 2] Ting Xu[1] Guozhong Qu[1] Xinhui Tu[1]
[1] Department of Computer Science, Huazhong Normal University 430079 Wuhan, China
[2] Software College of Tsinghua University 102201 Beijing, China
tthe@mail.ccnu.edu.cn egg622@ mails.ccnu.edu.cn qu_g_z@mails.ccnu.edu.cn tuxinhui@mails.ccnu.edu.cn

Abstract. How to improve the rankings of the relevant documents plays a key role in information retrieval. In this paper, a re-ranking approach based on topic words pair is proposed to improve precision while recall is preserved. The topic word pairs contain two correlated words, one of which is the original query word and the other come from the documents. The selection is based on Probabilistic Latent Semantic Indexing (PLSI). Then, the distribution of the word pairs is used to re-rank documents. Results show a 53.6% and 56.8% improvement compare to the initial retrieval without any re-ranking or query expansion on NTCIR-5 document collection for SLIR.

1 Introduction

Since the number of electronic Chinese documents is growing very fast, efficient techniques for Chinese information retrieval are needed. In the past years, significant progress has been made on this problem by researchers in the field of information retrieval (IR) and many retrieval models, indexing strategies, query expansion and re-ranking strategies have been successfully used in IR.

How to improve the rankings of the relevant documents plays a key role in information retrieval. In general, there are three kinds of information could be used for re-ranking:

1. Document information
   The clustering information has been used by Kyung-Soon Lee et al. [6]. Both the clusters and initial retrieval results have been used for re-ranking. Jaroslaw Balinski and Czeslaow Danilowicz [9] investigated ways to re-rank documents utilizing the inner distance between them.
2. Query information
   In the re-ranking phase, Mitra et al. [7] considered the correlation of query words. They add constraints to the query terms in order to get stronger indications for document relevance. Qu Youli et al. [8] described their method utilizing the location information of query terms.
3. Additional information
   Dequan Zheng et al.[10] adopt ontology information to re-rank documents.

Currently, re-ranking using query information alone has some difficulties: this method neglects query words' semantics. For example, document may be considered not relevant to query if they have not matching terms even though they are the same in semantics. Re-ranking using additional information (such as ontology, dictionary and so on) has other problem: the construction of ontology is still difficult. On the other hand, adding additional information into the query is equivalent to query expansion, which may cause a drift in the focus of the search topic. For example, 查询运动媒体或运动相关产业表彰老虎伍兹为运动明星的报导(Find documents about sports media or related enterprises recognizing Tiger Woods as a sports star), many system will add "高尔夫"(golf) to the query, as a result documents will get high score if they contain "高尔夫" with high frequency, though the document talk about other linksman.

Then in this paper, we present a re-ranking method utilizing both document information and query information. The main process is that: 1. Initial retrieval 2. Select topic word pairs. The topic word pair

[1] The paper is supported by National Natural Science Foundation of China (NSFC), (Grant No.60496323; Grant No.60375016 Grant No.10071028;); Ministry of education of China, Research Project for Science and technology, (Grant No. 105117).

means two words which strongly represent a topic. It contains exactly one of the original query words w_q and the other word w_d is from documents. The reason why use topic word pair can improve precision is that the two words are highly correlated. They describe the topic more exactly. They can focus on the search topic and give stronger indications of relevance. Thus, only the documents contain both w_q and w_d can be affected. 3. Re-rank documents based on the distribution of the topic word pairs.

Our topic word pair selection process is based on Probabilistic Latent Semantic Indexing (PLSI). Specifically, we look for the pairs with the strongest correlation in latent semantic space. The selection process is achieved automatically.

The structure of this paper is as follows. In the next section, we will give a brief review of hybrid indexing and retrieval model. In section 3, we describe the topic word pair and its selection. In section 4, we give the document re-ranking method. In section 5, we evaluate the performance of our proposed method and give out some result analysis. Finally, we conclude.

2 Retrieval Model and Indexing Strategy

Research work compared the retrieval effectiveness using different types of terms. In general, retrieval based on Chinese characters has the best recall while retrieval based on words or based on bi-grams has the best precision. To overcome the shortcomings of one type of term, we use hybrid index [1]. Firstly, words can be extracted automatically and then bi-grams are extracted at parts of the documents where the out-of-vocabulary problem occurs. Note that we also index single-character words if it is not a stop word. All these words and bi-grams used as index units. In addition, TFIDF retrieval model is used in our experiment.

3 Topic word pair

The topic word pair is two words which strongly represent a topic. It contains exactly one of the original query words and the other word is from documents. Let q = { $q_1, q_2 ... q_k$ } is the original query and D= $\{d_1, d_2, ..., d_n\}$ be a set of documents returned in response to the query by the initial retrieval. Let L be a set of word pairs, each one contains exactly one of the original query words [2],

$$L = \{ (w_i , w_j , \text{association\_intensity}) \mid w_i \in q \text{ and } w_j \notin q \}.$$

We extract L from the 1000 top-ranked documents in D and choose the "best" pairs from L, association_intensity is calculated according to the two words' correlation. This process is based on Probabilistic Latent Semantic Indexing [3]. Then, the pairs rather than single word will be used to the re-rank the documents. So such a method will only affects the score of documents which contains both w_j and w_j. We next describe the method in detail.

3.1 Construction of semantic space

3.1.1 Initialization

The index units are weighted by the word occurrence frequency and by the inverse document frequency as in Equation 5:

$$m(d, w) = tf(d, w) \times idf(w) \qquad (1)$$

Secondly, initialize $p(z_1)= p(z_2)=\ldots= p(z_n)=\frac{1}{n}$, where z_i is a latent class.Then associates an unobserved class variable z with each occurrence of a word w in a document d $\in$ r, thus, initialize P (z | d) and P (w | z) under constraint:

$\sum_{j=1}^{M} P(w_j \mid z_k) = 1$ and $\sum_{k=1}^{K} P(z_k \mid d_i) = 1$

3.1.2 Model Fitting with EM

E-step:

$$\mathrm{P\ (z \mid w,\ d)} = \frac{\mathrm{P(z)P(w \mid z)P(d \mid z)}}{\sum_{z' \in Z} \mathrm{P(z')P(w \mid z')P(d \mid z')}} \qquad (2)$$

M-step:

$$\mathrm{P\ (w \mid z)} = \frac{\sum_{d} m(d,w)P(z \mid d,w)}{\sum_{d,w'} m(d,w')P(z \mid d,w')} \qquad (3)$$

$$\mathrm{P\ (d \mid z)} = \frac{\sum_{w} m(d,w)P(z \mid d,w)}{\sum_{d',w} m(d',w)P(z \mid d',w)} \qquad (4)$$

$$\mathrm{P\ (z)} = \frac{\sum_{d,w} m(d,w)P(z \mid d,w)}{\sum_{d,w} m(d,w)} \qquad (5)$$

Following the likelihood principle, alternate (2)-(5) approaches a local maximum of log-likelihood in Equation 6:

$$\ell = \sum_{d \in D} \sum_{w \in W} m(d,w) \log P(d,w) \qquad (6)$$

3.2 Choose the best topic word pairs

We choose the best topic word pairs according to the weight of the pairs. In this approach the weight of the word pair is calculated utilizing the results above, the following formula was used:

$$\mathrm{P_{ij}} = \mathrm{Cos}(W_i\ ,\ W_j) = \frac{\sum_{z \in Z} p(w_i \mid z) p(w_j \mid z)}{\sqrt{\sum_{z \in Z} p(w_i \mid z)^2} \sqrt{\sum_{z \in Z} p(w_j \mid z)^2}} \qquad (7)$$

All pairs (w_i , w_j) containing exactly one of the original query terms are ranked by the P_{ij}. The top ranked pairs is then used to form a list for the original query. The list will be the base of re-ranking in the following step.

4. Document Re-ranking

Documents re-ranking is a method to sort the initial retrieved documents without doing a second retrieval. It's expected that the more a document is relevant the higher ranking it should have.

In this section we describe our re-ranking method which is based on distribution of topic word pairs. Our proposed approach integrates the information from the span of the two words in a pair, relative document frequency and the document-query similarity. As already defined, let D be the result set for query q, L be the list of topic word pair, *a* be a word pair in L. We re-rank D considering the following factors:

1. Span: For each word pair (w_i , w_j) in the list L, we consider the span

| position(w_i)-position(w_j) |

in document where position(w) is the sequential position number of w in text [4]. Intuitively, the shorter the span is, the higher is the weight of the document.

2. The weight of the word pair (w_i , w_j): The weight is calculated by formula 7, the higher it is, the closer w_i is related to w_j, thus the more important it is.

3. Relative document frequency: The ratio of document frequency of a word pair in the top 1000 document against the document frequency of it in the whole document collection. Obviously, the higher the ratio is, the more important the word pair is [5].

4. The original similarity S between document and query:

With both these information taken into consideration, document score is then calculated by the following formula.

$$\text{DocumentScore(d)} = \left(\sum_{a\in L}\frac{P_a \times df(a,D)/1000}{span(a)\times Df(a,C)/|C|}+1\right)\times S \quad \textbf{(8)}$$

Where d is a document, P_a is the weight of word pair a, span(*a*) is the span of the two words in *a*, df(*a* ,D) and Df(*a*, C) are the document frequency of a in D and whole document collection respectively, S is the original similarity between query and d.

Then, we use the new ranking score to re-order the 1000 documents.

5 Experiments and performance evaluation

5.1 Test collection

We evaluated re-ranking method with a test collection which is composed of NTCIR5 test collection for the single language (Chinese) information retrieval task. Then we only use field DESC as query and evaluate our methods on 20 topics. Standard trec_eval tool is used to compute the mean average precision (MAP) scores and the precision after R (= num_rel for a query) docs retrieved.

5.2 Experiments and results

In this section we report results from experiments we conducted to evaluate the influence of the suggested re-ranking method on search precision. As a baseline we use the lemur version 4.1 system.

The first experiment is designed to test how many topic word pairs used to re-rank documents affects the precision. Firstly we suppose that the number depends on the length of original query because too few topic word pairs have not enough information while too many bring much noise. Then (2m-1) word pairs are automatically chosen for each query empirically, m is the number of indexing unit in query.

The results of evaluation are presented in table 1. Column 2 (Rigid) displays the precision of rigid relevant measure and Column 3 (Relax) uses relax relevant measure. Column [normal] displays the precision of normal retrieval and column [Enhanced] displays the precision of using our proposed approach.

Table 1. Precision as a function of the length of query.

Length of query	Rigid		Relax	
m	normal	enhanced	normal	enhanced
m=3	0.2210	0.3112	0.2415	0.3781
m=4	0.2018	0.3247	0.2387	0.3725

m=5	0.2154	0.3031	0.2328	0.3547

From these results it is apparent that in general there is a significant improvement in precision when adapting our re-ranking method.

Table 2 shows the average precision improvement at R documents respectively. We can see that compared with original result without re-ranking, our proposed method can improve PreAt5 by 15% from 0.4960 to 0.5720 in relax relevant measure and improve 13% from 0.3520 to 0.3980 in rigid relevant measure.

Table 2. Precision at R documents.

Precision at R documents	relax		rigid	
	Original	After re-ranking	Original	After re-ranking
At 5 docs	0.4960	0.5720	0.3520	0.3980
At 10 docs	0.4660	0.5200	0.3420	0.3680
At 15 docs	0.4587	0.4953	0.3293	0.3487
At 20 docs	0.4360	0.4660	0.3150	0.3370

In the second experiment, we compared different topic word pair selection processes. The result is shown in Table 3. Column[Rigid] shows the average results of rigid relevant measure and Column[Relax] use the Relax measure. Row[PLSI] give the MAP result when PLSI is used to select topic word pair and Row[MI] select word pair based on mutual information. Both the selection method improves MAP clearly. And it finds out that the two selection processes achieve similar result, in the other words, the affections caused by topic word pair are not rely on specific arithmetic.

Table 3. measure for different selection methods.

different topic word pair selection	Rigid		Relax	
	MAP	%Change	MAP	%Change
Baseline	0.2024	-	0.2407	-
PLSI	0.3110	53.6%	0.3750	55.8%
MI	0.2980	47.2%	0.3772	56.7%

Finally, we present some of the queries and the topic word pairs used for re-ranking the documents in table 4.

Table 4. Some topic and their topic word pairs.

Query No.	topic	Some of its topic word pairs
006	<DESC>查询俄罗斯核子潜艇科斯克号意外沉没及等待救援的相关报导。</DESC> Find reports about the sinking and wait for rescue of the Russian nuclear submarine, Kursk.	科斯克 海军(kursk, navy) 潜艇 下沉(submarine, sinking) 潜艇 海底(submarine, benthal) 潜艇 丧生(submarine, death)
034	<DESC>查询美国追捕反美恐怖份子首领宾拉登的策略。</DESC> Find documents on U.S.A.'s strategy on pursuing Bin Laden, the leader of anti-American terror.	恐怖,奥玛(terror, Omar) 宾拉登,阿富汗(Ben Laden, Afghanistan) 宾拉登,搜捕(Ben Laden, manhunt) 恐怖，联盟(terror, league)

In our experiment, we get poor result on some query topics, to demonstrate the problem, we list topic 001 as following:

```
<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>CH</TLANG>
<TITLE>时代华纳，美国线上，合并案，後续影响</TITLE>
<DESC> 查询时代华纳与美国线上合并案的後续影响。</DESC>
</TOPIC>
```

We found the problem is most probably caused by the wrong segmentation. For example, 时代华纳 (Time Warner) will be segmented as 时代(time) /华纳(Warner) in which case 时代(time) will leads to the wrong meaning time, not a part of a company name. Then if we use word pairs containing 时代 (time) to re-rank documents, the documents talking about times will get high score. Such problem may be solved by using term as index unit. 时代华纳(Time Warner) can be treated as a whole company name in this way.

6 Conclusions

In this paper, we have proposed a method for re-ranking using distribution of topic word pairs. We have presented evidence that our proposed method is one which can produce significant improvements over the method based on similarity search ranking alone.

In this paper, how to choose the number of topic word pair is something to consider and we have supposed that the number depends on the length of original query. The experiments show it works generally and also show that using topic word pairs is not always effective if the word is wrongly segmented and leads to another meaning. Maybe term should be the better unit to form a pair. Another conclusion is that the selection process of topic word pair is not unique, it doesn’t rely on specific arithmetic.

For the future work, we will try to improve the quantity of topic word pair and do some experiments for re-ranking utilizing term pair.

Reference

1. Tsang, T.F., R.W.P. Luk and K.F. Wong, Hybrid term indexing using words and bigrams, Proceedings of IRAL 1999, Academia Sinica, Taiwan, 112-117, 1999.
2. David Carmel, Eitan Farchi, Yael Petruschka, Aya Soffer, Automatic Query Refinement using Lexical Affinities with Maximal information Gain. In Proceedings of the ACM SIGIR’02 Conference, Tampere, Finland, August 11-15, 2002.
3. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI (1999).
4. Olga Vechtomova, Murat Karamuftuoglu. Approaches to High Accuracy Retrieval: Phrase-Based Search Experiments in the HARD track.
5. L.P. Yang, D.H. Ji. I2R at NTCIR5. Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan.
6. Kyung-Soon Lee, Young-Chan Park, Key-Sun Choi. Re-ranking model based on document clusters. Information Processing and Management, 37(2001).
7. M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998. Second Informational Joint Conference on Natural Language Processing, Korea, October 11-13, 2005.
8. Qu Youli, Xu Guowei, Wang Jun. Rerank Method Based on Individual Thesaurus. NTCIR Workshop 2 Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization, 2001, 5.

9. Jaroslaw Balinski, Czeslaow Danilowicz. Re-ranking method based on inter-document distances. Information Processing and Management, 41(2005)759-775.

10.Dequan Zheng, Tiejun Zhao, Sheng Li, Hao Yu. A Hybrid Chinese Language Model based on a Combination of Ontology with Statistical Method.