Rerank Method Based on Individual Thesaurus

Qu Youli, Xu Guowei, Wang Jun
FUJITSU R&D Center CO., LTD.
Room 5201,EverBright International Trust Mansion, No.11 Zhong Guan Chun South Street
Haidian District, Beijing, China 100081
{quyouli,guowei,jwang}@frdc-fujitsu.com.cn

Abstract

This paper mainly introduces the rerank method based on individual thesaurus. This method is used to rerank the initial document set produced by a fulltext search system. Aimed at NTCIR, the individual thesaurus is constructed for each topic to produce query vector with different methods. The effectiveness of the rerank method is tested on the NTCIRII ad hoc Chinese text data and it proved that this method could promote precision of information retrieval.

Keywords: NTCIR, VSM, thesaurus

1 Introduction

In recent years, electronic documents become more common with the rapid development of Internet. As a consequence, it has become important for users to be able to easily retrieve the information they want from large document database. However, the keyword-based search engine, which is common in document information retrieval, does not meet these requirements well. It can often deal with complex logical expressions of queries and present the search output in a disordered manner. The disorder makes a user lose interest to look over. However there are obvious limitations in retrieval effectiveness because which documents are valuable depends on the context of the query---for example, the education, interests and previous experiment of a user [1].

Search engines generally treat search query in isolation. The results for given query are identical, independent of the user, or the context in which the user made the query. Context information may be provided by the user in the form of keywords added to a query. However, providing context in this form is difficult and limited. It is feasible to use pre-defined context for promoting the performance of search engine.

For this reason, individual thesaurus expressed the context of a user is used to rerank search output to meet the user's requirement of precision. The documents that the user need are ranked on the top. For this aim, search output, as initial relevant document set, must be produced firstly.

For NTCIR II ad hoc Chinese information retrieval, individual thesaurus for every topic, as the context of query, is constructed to describe the topic.

This paper does some experiments on the construction of the individual thesaurus, how to produce query with the individual thesaurus and how to set the term weight of query vector.

2 System architecture

System architecture (see figure 1) has two stages: stage of initial relevant document set and rerank stage. Stage of initial relevant document set: According to user's input of logic expression, an existing fulltext search system searches in document database and gets initial relevant document set. The existing fulltext search system that we used is Chinese search system (trial version) that is an adaptation of *IntelligentSearch* developed by FUJITSU LIMITED

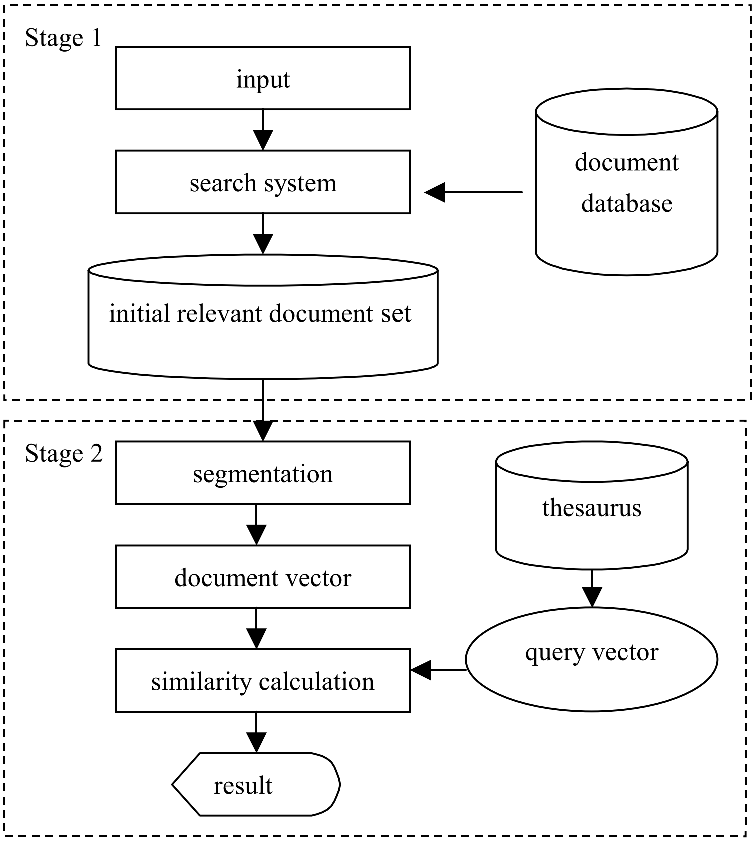


Figure1: system architecture

for Chinese document retrieval.

Rerank stage: Firstly, the documents in initial relevant document set are segmented; secondly, the word frequency statistics of the segmented documents is done and the documents vectors are gotten; last, the initial documents are reranked and output to a result file according to the similarity scores between the vector of the document in the initial relevant document set and the query vector which comes from the individual thesaurus. The followings are the illustration of the key problems in the figure 1.

3 Initial relevant document set

To meet the recall, the query condition to the Chinese search system is as low as possible. The input logic expressions are made out by manual. The main steps are as follows:

(1)Make out query terms by manual according to each topic of NTCIRII. Then the initial query expression comes out after the "and" operation is done between these terms.

(2)Make out the synonymous table of every term.

(3)The query terms are replaced by an expression which is gotten after the "or" operation is done between the query terms and their expanded synonymous terms.

The logic expression is input to the Chinese search system and the relevant document set is brought out. In this paper, there are 4455 documents to 50 topics.

4 Rerank

The aim of rerank to the initial relevant document set is to improve the query precision. In this paper VSM (vector space mode) is used to reach the aim. The method is to calculate the similarity between the query vector and every document vector. The greater the similarity is, the greater their relevance is. After the initial relevant documents are segmented by the NLP technique, the term of the query vector and document vector can be expressed by word.

4.1 Individual thesaurus

The element of the individual thesaurus is the pair of (lemma, attribute_set), where lemma is a word and attribute_set is the set of the word attributes. The latter shows the lemma belongs to which of the four fields: <title>, <question>, <narrative>, <concept>. Different fields have different weights and the term weight of query vector will be calculated based on it.

The method to construct individual thesaurus is as follows:

(1)Get out lemma from concept field in turns;

(2)Record the information of the field of the lemma or its relevant word in the attribute_set;

(3)Repeat (1) and (2) until all the lemma of concept field are done.

In individual thesaurus, “t” means title field, “q” means query field, “n ” means narrative field, and “c” means concept field. for example:“科省 tqnc” means “科省” or one of its relevant words belongs to t,q,n,c fields.

The following example is about the individual thesaurus of topic06.

科省	tqnc	柯省	tqnc
科索沃	tqnc	柯索伏	tqnc
难民	tqnc	难民潮	tqnc
难民营	tqnc	援助	qnc
收容	qnc	救援	qnc
医疗	qnc	马其顿	nc
土耳其	nc	外交部	c
国际	c	联合国	c
红十字会	c		

4.2 Segmentation

One of the discriminations between Chinese documents and English documents is that the former has no laps between words. So the first step in Chinese language process is segmentation. If the segmentation has a higher precision, the document vector whose term is word can describe the thing more clearly.

The segmentation tool that we used is an adaptation of *Breakfast* developed by FUJITSU LABORATORIES LTD. for Chinese[2][3]. Hidden Markov Mode is employed in this tool. In sentence segmentation, all the possible alternative ways will be attempted. With the word frequencies in a lexicon and connection probabilities between POSes in syntactic rules, the probability of each way can be gotten. The first choice is the way with maximum probability. The information about word frequency and POS used in above-mentioned computation comes from *ACADMIA SINICA BALANCED CORPUS* [4] by statistical method.

4.3 Similarity calculation

VSM is employed to calculate the similarity between query vector and document vector. The term of vector is word. If $T=\{t_j\}$ is term set, $|T|=n$ （$|T|$ shows the size of set T）,then query vector v_j of topic j can be express $v_j=(v_{j1},v_{j2},\cdots,v_{jn})$, in which v_{jk} denotes the weight of t_k in v_j. The vector $D_i=(d_{i1},d_{i2},\cdots,d_{in})$ denotes a document, d_{ik} denotes the weight of t_k in D_i. The similarity between v_j and D_i is calculated by following formula.

$$s_j = \sum_{k=1}^{n} d_{ik} \times v_{jk}$$

4.3.1 Query vector

The terms of query vector all come from the individual thesaurus corresponding to the topic. v_{jk}, the weight of term k in query vector j, depends on the attribute_set of the term k. We use two methods to set v_{jk}.

Method A: Set v_{jk} with the size of the attribute_set

$$v_{jk} = f(|attribute\_set|)$$

$|attribute\_set|$ means the size of attribute_set.

Method B: Set v_{jk} with the element of the attribute_set

$$v_{jk} = \sum_f fw(f)$$

f is the topic field (title, question, narrative, concept); fw is weight of the field.

4.3.2 Document term weight

In document vector, document term weight d_{ij} is calculated with a modified tf part of OKAPI[5]:

$$d_{ij} = \frac{10 * term\_freq}{4.5 + \frac{b * dl}{adl}}$$

dl : document length in byte.

adl : average document length in byte.

If a term presents in the document title, the term is more important. In order to express this idea, we modify $term\_freq$ part.

$$term\_freq = content\_freq + title\_occur * \alpha$$

$content\_freq$ denotes the times that the term presents in the document text.

$title\_occur$ denotes whether the term presents in the document title. If the term presents in the document title, then $title\_occur = 1$, otherwise, $title\_occur = 0$.

α denotes the ratio of the importance in the document title to the document text. α is 10 in our experiments.

5 Experiments

The number of terms in a query vector influences the performance of rerank. How the number of terms influences the performance of rerank and how many terms is more appropriate are to be decided by experiments. At the same time, the weight of terms influences the performance of rerank as well, and how to set weight is also important. Followings are the experiments dealing with different numbers of terms and various weights respectively.

5.1 Experiments 1

Because the terms of query vector used in the rerank stage come from the individual thesaurus, the choice lemma as a term of query vector influences the rerank performance. According to the choice rule of lemma in the individual thesaurus, three experiments, tq, tqn, tnc, is designed and is compared with initial relevant document set. The term weight of query vector in those experiments is defined as the formula described in method B of 4.3.1. In the formula, fw(t)=3,fw(q)=1.5,fw(n)=0.5,fw(c)=0.5.

tq: the experiment that is selected the lemma whose attribute set includes t or q, the average number of terms in the query vector is 8 or so;

tqn: the experiment that is selected the lemma whose attribute set includes t, q or n, the average number of terms in the query vector is 14 or so;

tqnc: the experiment that is selected the lemma whose attribute set includes t, q, n or c, the average number of terms in the query vector is 20 or so;

initial: the pre_reranked result.

NTCIR provides two correct sets: the rigid correct set and the relax correct set. The two correct sets are used to evaluate the experiments. The evaluation result is as table 1(average precision table) and figure 2. The rigid R-P curve in figure 2 means recall-precision curves for the average of 50 topics by the rigid correct set. The relax R-P curve in figure 2 means recall-precision curves for the average of 50 topics by the relax correct set.

Table 1 : average precision table

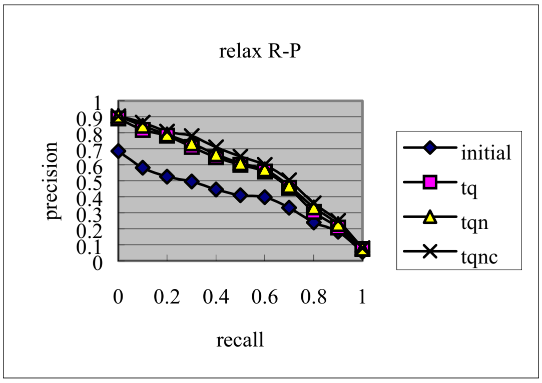
	initial	tq	tqn	tqnc
rigid correct set	0.247	0.442	0.447	0.488
relax correct set	0.356	0.536	0.551	0.579

Conclusions:

(1) The method of adopting individual thesaurus can improve retrieval precision. Under rigid correct

set, tqnc increases 24 percent point compared with initial result;

(2) In the present point, the more items in the query vector lead to better result. But an optimum has not been found about how many is appropriate. Through this group of experiments, it can be seen that using a smaller thesaurus can improve retrieval precision to a much degree. In addition, how to extend related words is a difficult problem. In the future work on how to extend related words will be carried out.



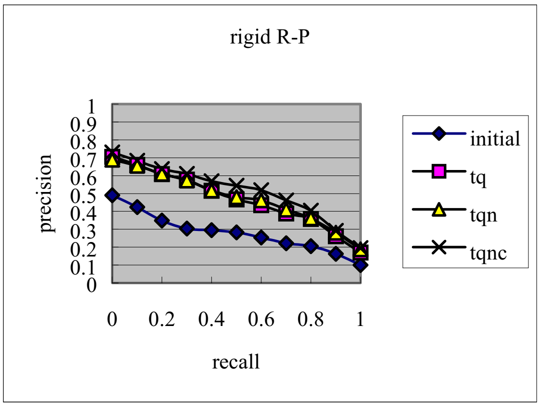


Figure 2: recall-precision curves

5.2 Experiments 2

Experiments 2 is about how to set weight to terms of query vector. We do some experiments with method A and method B in section 3.3.1 respectively.

5.2.1 Set term weight of the query vector with method A

Because VSM is used to calculate similarity, the similarity score is only related to ratio among term weights if setting term weight with method A. The different ratios influence the performance of precision. So we design a serial of experiments for the different ratio. It is supposed f(2)=f(1). The evaluation result of the experiments shows in following table 2.

AveP(rigid) is the mean of the average precision scores of each of the individual topics under rigid correct set.

AveP(relax) is the mean of the average precision scores of each of the individual topics under relax correct set.

Table 2 : experiments with method A

No	f(4)/f(3)	f(3)/f(2)	AveP (rigid)	AveP (relax)
1	0	1	0.3011	0.4141
2	0.2	1	0.4308	0.5404
3	0.5	1	0.4496	0.5529
4	1	1	0.4560	0.5586
5	2	1	0.4796	0.5735
6	3	1	0.4817	0.5746
7	4	1	0.4808	0.5740
8	5	1	0.4790	0.5736
9	6	1	0.4771	0.5736
10	7	1	0.4776	0.5740
11	8	1	0.4777	0.5746

In those experiments, the performance of experiment 6 is the best and experiment 1,2,3 is worse. The reason is that f(4)/f(3) is appropriate in experiment 6 but they are less in experiment 1,2,3. If f(4)/f(3) is less, the influence of term whose attribute _set is tqnc is less. But this kind of term is the most important in topics.

We can draw a conclusion that the performance of rerank is better when f(4)/f(3)>=2. When f(4)/f(3)=3, the average precision is 0.4817. With the gradual increase of f(4)/f(3), the average precision falls slowly. This is because the function of the terms including 3 attributes trails off with the weight of the term including 4 attributes increases.

5.2.2 Set term weight of the query vector with method B

The different ratio among field weights influences the performance of precision if setting set term weight with method B. So we designed a serial of experiments. The evaluation result of the experiments shows in following table 3.

fw(n)/fw(c) is set to 1, AveP(rigid) and AveP(relax) in table 3 as defined in 5.2.1.

Table 3: experiments with method B

No	fw(t)/ fw(q)	fw(q)/ fw(n)	AveP (rigid)	AveP (relax)
1	0.5	3	0.4759	0.5710
2	1	3	0.4778	0.5725
3	2	3	0.4888	0.5796
4	3	3	0.4805	0.5735
5	4	3	0.4805	0.5738
6	5	3	0.4790	0.5732
7	2	2	0.4806	0.5729
8	2	4	0.4814	0.5735
9	2	5	0.4813	0.5736
10	1	1	0.4787	0.5723
11	8	1	0.4661	0.5613

In these experiments, the performance of experiment 3 is best. With gradual increase of fw(t)/fw(q) from 0.5 to 2, AveP increases gradually. With gradual increase of fw(t)/fw(q) from 2 to 5,AveP decreases gradually. This demonstrates that the value of fw(t)/fw(q) should be within measure.

The different ratios among field weights influence slightly the performance of precision. The reasons are that the lemmas having attribute t often have attribute q,n,c and the lemmas having attribute q often have attribute n and c. So the change of the ratios among field weights influences slightly the ratio among term weights.

6 Conclusions

We presented a rerank method based on individual thesaurus for NTCIRII Chinese ad hoc retrieval task. We proposed and experimented with various techniques and parameters.

Tested on the NTCIRII Chinese ad hoc text data, the method of adopting individual thesaurus can improve retrieval precision. The more items in the query vector lead to the better result. The term of query vector is most important if it has the attribute t. Method B setting the term weight can reach a better performance than Method A. Both method B and method A can't reach more precision because they all take a document as bag of words. Refinements of proposed methods are in progress to further improve the performance of the current retrieval system.

Acknowledgement

The authors would like to thank Institute of Information Science Academia Sinica for *ACADMIA SINICA BALANCED CORPUS*.

References

[1]Steve Lawrence. Context in Web Search, IEEE Data Engineering Bulletin, Volume 23, Number 3,pp.25-32, 2000

[2]段慧明,松井久仁於,徐国伟,胡国昕,俞士汶.大规模汉语标注语料库的制作与使用,语言文字应用,pp72-77,200年第2期.

[3]飒々野学,斎藤由香梨,松井くにお.アプリケーションのための日本語形態素解析システム.言语处理学会第3回年次大会发表论文集, C4-7, pp.441-444, 1997年

[4]Chinese Knowledge processing group, Institute of Information Science Academia Sinica ,the content and explanation of *ACADMIA SINICA BALAANCED CORPUS*, 1995.9

[5]S E Robertson, S Walker and M Beaulieu. Okapi at trec-7. The seventh Text REtrieval Conference, 1999