Question Analysis and Query Expansion in CS-CS IR4QA

Maofu Liu, Fang Fang, Qing Hu, Jianxun Chen
College of Computer Science and Technology
Wuhan University of Science and Technology
liumaofu@wust.edu.cn, fangfang0402@126.com

Abstract

This paper describes our work in NTCIR-7 on the subtask of simplified Chinese monolingual information retrieval for question answering (CS-CS IR4QA). Based on the observation that inappropriate key terms and term mismatch often result in depressed precision and impressive recall, we employ a special question analysis method extracting more appropriate key terms and apply the query expansion technique gaining more relevant key terms, to enhance precision and efficiency for retrieval performance.

Keywords: *term mismatch, question analysis, query expansion.*

1. Introduction

There are many important factors that affect the performance of information retrieval. The key words extracted for query make significant influence on the retrieval performance[1]. The key terms extracted from a question in IR4QA can be different with distinct segmentation strategies, because one long term in a question may be segmented as only one term or more short terms. The other major problem in information retrieval is term mismatch between query and documents[2]. The problem of term mismatch in information retrieval often occurs because the expression of the same concept may be represented by different terms between the authors' documents and users' query statements. These two factors may lead to information lost and information overload. As a result, the retrieval system may get low rate of recall and precision[3].

The query expansion is an effective way to solve term mismatch problem by expanding the key terms with a certain number of other related terms in the initial query. The most well-known query expansion technique is pseudo-relevance feedback (PRF)[4] and it is widely used in the open information retrieval evaluation workshops for improving the retrieval effectiveness. And for many times, TREC evaluation has demonstrated that PRF is a simple but very effective query expansion technique. Firstly, several documents are retrieved as a result of the initial retrieval. It holds an assumption that the top-n retrieved documents are relevant, the system looks on the terms contained in the initial retrieval result as expansion terms and implements the second retrieval based on them.

In order to extract appropriate key terms from the original question, we use a specific question analysis approach to get more appropriate key terms for query. Then, we adopt the PRF technique and co-occurrence based and metric correlation based query expansion approaches to reduce the negative impact of term mismatch on retrieval performance.

The remainder of this paper is organized as follows. Section 2 delineates system architecture. Section 3 describes the question analysis technique in detail. Section 4 elaborates on the query expansion. Section 5 discusses our evaluation results. Finally, we conclude our paper in section 6.

2. System architecture

Our system includes four main modules, i.e. indexing processing, question analysis, retrieval and query expansion. Figure 1 illustrates our CS-CS IR4QA system architecture in detail.

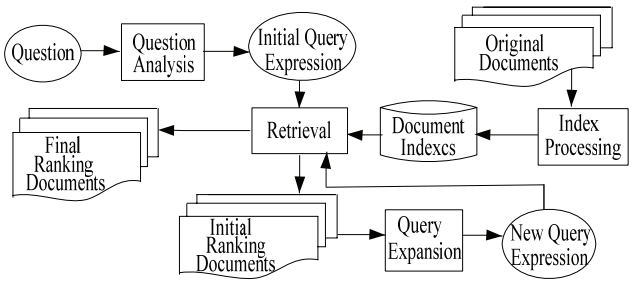

Figure 1. System architecture

(1) Index processing

The index processing module processes the original documents firstly. The Chinese words can not be segmented by obvious separators in document, so we do the word segmentation by applying a maximum and

reverse maximum match algorithm based on a dictionary[5]. The dictionary collects most of the common words, such as base words, person names, location, and so on. After segmentation, each word will be an index unit stored in an inverted index file to construct document indexes.

(2) Question analysis

When the user gives a question, we should make question analysis to extract key terms from the question to form the query expression for the initial retrieval.

(3) Retrieval

Our system employs the Vector Space Model (VSM) to determine the relevance between the given documents and the user query[6].

There are two retrieval phases in our system. The initial ranking documents can be gained by the initial retrieval phase using the initial query expression. After query expansion introduced in the section 4 below, the second retrieval results will be obtained using the new expanded query expression.

(4) Query expansion

When we get the initial ranking documents, the query expansion module extract terms from the top-n initial ranking documents, and then the new terms will be added to the initial query expression to construct a new one for the second retrieval phase.

3. Question analysis

The purpose of question analysis is to extract key terms from the original question to make up an initial query expression for the initial retrieval.

When the user gives a question, the system will segment it into words based on the same dictionary in the index processing module. But there may be more than one choice for the question segmentation. For example, “诺贝尔奖” can be segmented as “诺贝尔奖” only as one term or “诺贝尔+奖” as two terms. We assume that long word match can increase scores of semantic information to generate more precise result to fit in human cognition. In our experiment, we found long word match indeed help improve MAP value. In some situations, there may be no the long term in the original documents because some other short terms may express the same idea or some other long terms contain the same short terms but in a different word sequence[7]. As a result, few documents can be retrieved and a large number of documents are ignored. To solve this problem, we use the question analysis technique illustrated in Figure 2.

If some terms can be segmented as both short and long terms at the same time, we choose the longest ones as the key terms and retrieve the original documents at first. If the returned documents number *r* is less than *s*, which denotes the threshold value when use the key terms to search the relevant documents, the key terms can be segmented into the shorter ones and retrieve the original document again. Otherwise, i.e., the *r* is not less than *s* the system will implement the query expansion processing directly. One example is shown in Figure 3.

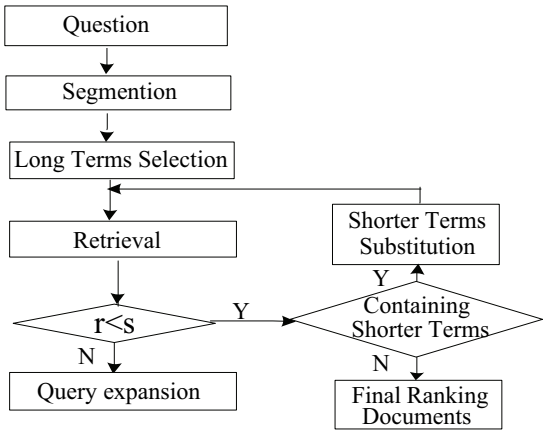

Figure 2. Question analysis process

As shown in Figure 3, when retrieving with the long terms selected above, the number of returned document is lest than *s* (the values of s is 100 in our experiment), so we substitute the term “亚洲金融危机” with “亚洲+金融危机” and retrieve again.

Query:列举亚洲金融危机对经济的影响。(List the impact of the Asian financial crisis on the economy.)
Segmentation:
列举,亚洲金融危机(亚洲,金融危机(金融,危机)),对,经济,的,影响.
Long term selection: 亚洲金融危机,经济,影响

Figure 3. Example of question analysis

4. Query expansion

In order to expand the query exactly, we should focus on extracted terms associated with the initial query terms. In this paper, we use two approaches to obtain these related terms.

4.1. Co-Occurrence based query expansion approach (CO)

This approach is derived from the same idea as LCA (Local context analysis)[8]. The co-occurrence means some words appear together in a “window”, and the “window” is defined as a document unit in our system. We try to seek the co-occurrence of key term *q* in the query and word *w* in the top *n* (*n* is assigned to 50 in our

experiment) documents from the initial retrieval. The co-occurrence degree of query term and word from each document d of the top n documents, called $cohd$, is calculated by the following equations[9].

$$cohd(w,q \mid d) = \prod_{q \in Q} (cood(w,q \mid S) + 1.0) \quad (1)$$

$$cood(w,q \mid S) = \frac{\sum_{d \in S} \log(tf(w \mid d)+1.0) \times \log(tf(q \mid d)+1.0)}{n} \quad (2)$$

Where S is the top-n documents set, Q is the query terms set, $tf(w \mid d)$ and $tf(q \mid d)$ denote the frequency of w and q in d respectively.

Some words may have high frequency but low distinction, we use the inverse document frequency logarithm to discriminate them and modify the equation as follows.

$$f(w,Q \mid C,S) = \sum_{q \in Q} idf(q \mid C) idf(w \mid C) \log(cood(w,q \mid S) + 1.0) \quad (3)$$

Where C indicates all of the documents set in the initial retrieval.

The system uses equation (3) to calculate the co-occurrence degree for each word and chooses the top *m* words as expanded terms.

4.2. Metric Correlation based query expansion approach (MC)

In this approach, we extract words for expansion based on the idea of metric correlation[10]. We define the distance between two terms as $r(t_u, t_v)$, if t_u and t_v are in the same document. We also employ equation (3) to calculate the scores of words, and the only difference is the equation of $cood(w,q \mid S)$, defined as follows.

$$cood(w,q \mid S) = \frac{\sum_{d \in S} \sum_{w \in d} \sum_{q \in d} \frac{1}{r(q,w)}}{n} \quad (4)$$

Where $r(q,w)$ is the position distance between q and w in d.

4.3. Weight of the expanded terms

When the terms for expansion are extracted, they are added to the initial query.

Obviously, different expanded terms have different significance, so different weight should be assigned to them. A common way is applying the Rocchio equation[11] directly to compute the weight $w(q \mid Q_{new})$ of each term in the new query Q_{new}. In the system, we modify this equation, because we consider the score for the expanded term evaluated by the equation (3) also indicates the importance of the terms. Then the equation is defined as follows.

$$w(q \mid Q_{new}) = p \cdot w(q \mid Q) + k \cdot avg(boost) \cdot \frac{score(q)}{MaxScore} w(q \mid d) \quad (5)$$

Where $w(q \mid Q)$ is the weight of key term q in the original query Q, $w(q \mid d)$ is the weight of q in document d, n is the number of top selected documents, and p and k are experimentally determined positive constants. $boost$ is one factor as a multiplier besides the factors tf and idf to compute the weight of the query key term in the initial query, and the $avg(boost)$ is the average value of them. The $score(q)$ indicates the score value computed by equation (3) for each expanded key term q and $MaxScore$ is the maximum of them.

In our experiment, p is assigned to 1.0, k is set 0.9 and 0.95 when apply co-occurrence and metric correlation based query expansion approaches respectively.

5. Experiments

We submitted five formal run files to NTCIR-7 and the official evaluation results of performance based on the pseudo-qrels and real-qrels[12] are listed in Table 1.

In Table 1, "Run" indicates the name of the run file we submitted which the prefix "NLPAI-CS-CS-" is omitted. The suffix "T" indicates the question title and "DN" is the description and the narrative of the question. The "Analysis File" means the question analysis files offered by other participators. "QE" indicates the query expansion approach. "PQ" and "RQ" denote the performance based on the pseudo-qrels and real-qrels respectively.

Table1. Formal run experiment official results

Run	Analysis File	QE	Mean AP		Mean Q		Mean nDCG	
			PQ	RQ	PQ	RQ	PQ	RQ
01-T	No	MC	0.2801	0.1198	0.2967	0.1099	0.4186	0.2383
02-T	No	CO	0.2615	0.1379	0.3349	0.1227	0.4743	0.2536
03-T	CMUJAV-CS-CS-01-T	CO	0.3010	0.1170	0.3010	0.1074	0.4395	0.2297
04-T	Apath-CS-CS-01-T	MC	0.2711	0.1117	0.2801	0.1014	0.4048	0.2204
05-DN	CSWHU-CS-CS-03-DN	CO	0.3261	0.1302	0.3261	0.1211	0.4613	0.2493

Comparing with other groups, our system does not achieve a good official result. We think the main reason is that we only submitted 10 documents for each question in the final retrieval result. But if we present more result documents, e.g. 1000 document IDs as the limit for each question, and evaluate the results with the IR evaluation package offered by NTCIR, the additional experiment results are better and they are listed in Table 2.

Table 2. Additional experiment results

Run	Mean AP	Mean Q	nDCG
01-T	0.4316	0.4510	0.6763
02-T	0.4682	0.4879	0.7051
03-T	0.4412	0.4452	0.6031
04-T	0.4241	0.4298	0.6139
05-DN	0.4720	0.4825	0.6724

In Table 2, the run files are same as the ones in Table 1, with the same question analysis files and query expansion approaches. When we pop at most 1000 document IDs for each question, the results are much better than the results in Table 1 which contain up to 10 document IDs.

We test the results which contain document IDs from 100 to 1000 maximally, we find that although the run files combining different question analysis file and query expansion approach, all of them get the same trend that the more documents IDs, the better result. Figure 4 shows the results of the run file 02-T using our own question analysis method and the CO query expansion approach.

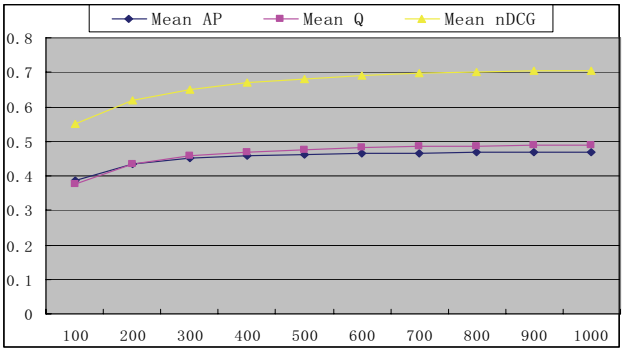

Figure 4. Results with different maximal document IDs

In Figure 4, X-axis denotes the document IDs limit number and "100" means the maximal document IDs is 100 for each question. Y-axis shows the values of the evaluation results.

From Figure 4, we can drive the conclusion that the experiment results based on the real-qrels are closely relative with the number of the retrieval result document for each question.

As the run file 03-T, 04-Tand 05-DN in Table 1 and 2, we use the question analysis files of other NTCIR groups, they are CSWHU, Apath and CMUJAV respectively. The Figure 5 below shows the detailed information of the key terms for the question topic "ACLIA1-CS-T384".

CSWHU-CS-CS-01-T:
<KEYTERMS>
<KEYTERM SCORE="1.0">宇宙大爆炸</KEYTERM>
<KEYTERM SCORE="0.3">理论</KEYTERM>
</KEYTERMS>
Apath-CS-CS-01-T:
<KEYTERMS>
<KEYTERM SCORE="1.0"> 宇 宙 大 爆 炸 理 论 </KEYTERM>
</KEYTERMS>
CMUJAV-CS-CS-01-T:
<KEYTERMS>
<KEYTERM SCORE="1.0">宇宙</KEYTERM>
<KEYTERM SCORE="1.0">大</KEYTERM>
<KEYTERM SCORE="1.0">爆炸</KEYTERM>
<KEYTERM SCORE="1.0">理论</KEYTERM>
<KEYTERM SCORE="1.0"> 宇 宙 大 爆 炸 理 论 </KEYTERM>
<KEYTERM SCORE="1.0"> 宇 宙 大 爆 炸 理 论 </KEYTERM>
<KEYTERM SCORE="1.0"> 宇 宙 大 爆 炸 </KEYTERM>
<KEYTERM SCORE="1.0">宇宙大爆炸</KEYTERM>
</KEYTERMS>

Figure 5. Key terms in question analysis files

From Figure 5, we simply gain the key term from the XML tag <KEYTERM> and assign the value of its attribute "SCORE" to the boost value. For example, "宇宙大爆炸" and "理论" are the key terms of the question topic "ACLIA1-CS-T384" in the CSWHU question analysis file and their boost values are "1.0" and "0.3" respectively. We can find that there is a little difference in CMUJAV comparing to CSWHU and Apath. Since "宇宙 大 爆炸 理论" and "宇宙 大 爆炸" are simple composition of four and three key terms appeared above lines and separated by space, then we omit the these KEYTERM lines. The evaluation results with the CO query expansion approach and 300 maximal result document IDs are listed in Table 3.

Table 3. Topic ACLIA1-CS-T384 analysis

Run	Key Terms & boost	AP	Q	nDCG
CSWHU	{(宇宙大爆炸:1.0), (理论:0.3)}	0.7720	0.8342	0.9387
Apath	{(宇宙大爆炸理论:1.0)	0.3987	0.3987	0.5908
CMU	{(宇宙:1.0),(大:1.0), (爆炸:1.0),(理论:1.0)}	0.4044	0.4044	0.5931
NLPAI	{(宇宙大爆炸:1.0) (理论:1.0)}	0.4079	0.4981	0.7847

From Table 3, we find that the results of CSWHU are obviously better than ours even though having the same key terms, i.e. “宇宙大爆炸” and “理论”, but the different boost values. If we assign the boost values of “宇宙大爆炸” and “理论” to 1.0 and 0.3 respectively, the retrieval system will emphasize on the information “宇宙大爆炸”.

Here we consider about the whole performance of the five run files in Table 1 and 2.As shown in the two tables, 02-T gains the best results, and 05-DN the second. In fact, there are only a few numbers of queries which cover different key terms extracted for the 02-T and 05-DN run files. So we mainly consider the effect of boost value set for the key terms. As mentioned in the above paragraph, different boost value set for key terms can make contribute to the retrieval performance. But we must regulate the boost value rationally, otherwise, it will fail.

The following Table 4 is another example about question topic ACLIA1-CS-T385 with the question title “列出与北京大学百年校庆相关的大事”.

Table 4. Topic ACLIA1-CS-T385 analysis

Run	Key Terms & boost	AP	Q	nDCG
CSWHU	{(北京大学:1.0), (百年:1.0), (校庆:0.3), (大事:0.3)}	0.0437	0.0475	0.2228
NLPAI	{(北京大学:1.0), (百年:1.0), (校庆:1.0), (大事:1.0)}	0.4283	0.4226	0.6642

Obviously, the key term “校庆” is important for the topic ACLIA1-CS-T385. In Table 4, if we set the boost value 0.3 for it which is much lower than the value 1.0 set for the key term”北京大学” and ”百年”, the retrieval system will mainly search the information ”北京大学” and ”百年”, and “校庆” will be ignored. Consequentially, the result will become even worse than the same boost value 1.0 set for all of the key terms.

By the comparative experiment, we can draw the conclusion that if the key terms are set proper scores for the boost values, it will improve the retrieval performance. Otherwise, it will decrease the performance. So the unreasonable boost value setting for the key terms is also one of the causes why the run file “05-DN” gets worse results than “02-T”.

Reviewing the Table 1 and 2, the result of “04-T” is the worst one. In fact, the scores of the key terms are also set as the same value 1.0, but almost all of its terms are longer ones, such as “世界各地”, “宇宙大爆炸理论”, and so on. Through our question analysis, we find these long terms are not applicable and re-segment them into shorter ones “世界+各地” and “宇宙大爆炸+理论” respectively. Then we can get a better retrieval performance. This indicates that our question analysis approach is really effective when long terms are not appropriate key terms for retrieval. We can also find that the co-occurrence based query expansion approach works better than the metric correlation in our system.

6. Conclusions

In order to solve the problems of inappropriate key terms exacted from the initial question and term mismatch, we presents a question analysis approach to exact appropriate key terms and apply query expansion technique to get more useful key terms for the query. We also combine these techniques with the word-unit based index files and VSM information retrieval model to check their effectiveness. By experiments, we find that these techniques can reduce the impact of the term mismatch and enhance the retrieval performance.

Through comparative experiments, we demonstrate that the evaluation results are apparently increased if we give more query result relevant documents. We also find that if the key terms are set proper scores for the boost values, the retrieval performance will be significantly improved.

Acknowledgments

The work in this paper was supported partially by a Special Research Grant from Wuhan University of Science and Technology (No. 50003301) and partially by Research Grant from Hubei Provincial Department of Education (No. 500064).

References

[1] Pante P. and Lin D. A statistical corpus-based term extractor. Proceedings of AI 2001[C]. Ottawa, Canada, Springer-Verlag, 2001, 36-46.

[2] Salton G. and McGill M. An Introduction to Modern Information Retrieval, New York, NY: McGraw-Hill. 1983.

[3] Rijsbergen van. A new theoretical framework for information retrieval[C]. In Proceedings of 1986 ACM Confefence on Research and Deve1opment in Information Retrieval, 1986, 194-200.

[4] Buckley. C, and Salton. G, J. Allan. Automatic retrieval with locality information using Smart. Proceedings of TREC 1, 1992, 59-72.

[5] Miao Douqian and Wei Zhihua. The principle and application of Chinese text information process. Tsinghua University Press, 2007, 22-23.

[6] Ricardo B.-Y., Berthier R.-N., et al. Modern Information Retrieval. China Machine Press[C] 2006, 20-24.

[7] Ekmekcio F. Effectiveness of query expansion in ranked-output document retrieval systems. Journal of Information Service, 1992, 18(2):139-147.

[8] Xu J. X. and Croft W. B. Improving the Effectiveness of Information Retrieval with Local Context Analysis[J].ACM Transactions on Information Systems, 2000, 18(1):79-112.

[9] Ding G., Bai S. and Wang B.. Local Co-occurrence based Query Expansion for Information Retrieval. Journal of Chinese Information Processing, 2006, 20(3):84-91.
[10] Ricardo B.-Y., Berthier R.-N., et al. Modern Information Retrieval. China Machine Press[C] 2006, 89.
[11] Rocchio J. Relevance feedback in information retrieval[A]. The Smart Retrieval System--Experiments in Automatic Document Processing[M], 1971, 313-323.
[12] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Ji, D., Chen, K.-H., Nyberg, E. Overview of the NTCIR-7 ACLIA IR4QA task, Proceedings of NTCIR-7, to appear, 2008.