Wikipedia Article Content Based Query Expansion in IR4QA System

Maofu Liu, Bin Zhou, Liwen Qi and Zilou Zhang
College of Computer Science and Technology
Wuhan University of Science and Technology
liumaofu@wust.edu.cn, zb_zhoubin@163.com

ABSTRACT

This paper describes the work of our WUST group in NTCIR-8 on the subtask of English to Simplified Chinese and Simplified Chinese to Simplified Chinese information retrieval for question answering (EN-CS and CS-CS IR4QA). In order to enhance the precision and efficiency in question analysis, we employ a special question analysis method extracting more appropriate key terms and apply the query expansion technique gaining more relevant key terms based on Wikipedia article content related to the query.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval – *Retrieval models*.

General Terms

Experimentation

Keywords

IR4QA, Wikipedia Article, Query Expansion

1. INTRODUCTION

The NTCIR Workshop is a series of evaluation workshops designed to enhance research in Information Access (IA) technologies including information retrieval, question answering, text summarization, information extraction, etc. Our subtask in NTCIR-8 is English to Simplified Chineses and Simplified Chineses to Simplified Chineses IR4QA from ACLIA task and the ACLIA task is to accept complex questions and factoid questions and provide answers in Chinese (Simplified, Traditional) and/or Japanese. The target corpus consists of newspaper articles.

In information retrieval system, users often submit the query which is a short description by natural language, and they decides the relevance of document not based on semantics of query terms in documents, but existence of query terms. If the IR system just simply checks the existence of query terms in documents without taking the context of documents into account, it often causes term mismatch and declines the performance greatly [1]. The other major problem in information retrieval is the key words extracted for query [2]. 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. 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. By our work in NTCIR-7, we find that this method can indeed reduce the impact of the term mismatch and enhance the retrieval performance. Therefore, we extend this approach, and try to make query expansion based on the related Wikipedia article content.

As we know, Wikipedia is a multilingual, web-based, free-content encyclopedia project, and each of its article provides information to explain the term of the article title. In order to expand the most relevant terms, we make use of related Wikipedia article as the “seed” document of the related question. And then make question expansion based on the most relevant paragraph of the question.

The remainder of this paper is organized as follows. Section 2 delineates system architecture. Section 3 describes the query

expansion via Wikipedia article content in detail. Section 4 discusses our evaluation results. Finally, we conclude our paper in section 5.

2. SYSTEM ARCHITECTURE

Our system includes four main modules, i.e. indexing processing, question analysis, query expansion and retrieval. Compared with the original system, we have introduced a query expansion in external resources mainly from Wikipedia, some from Baidu and Google. Figure 1 illustrates our IR4QA system architecture in detail.

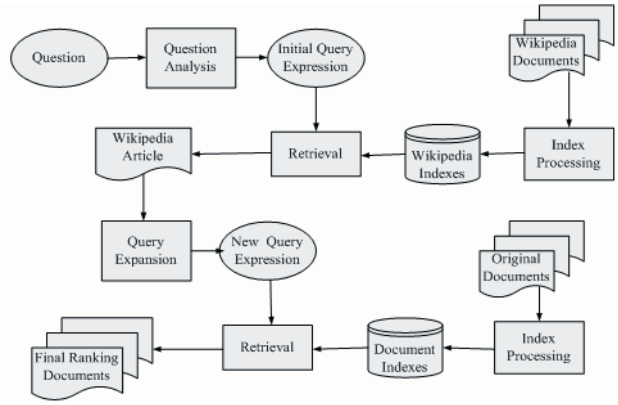

Figure 1. System architecture

(1) Indexing processing

The index processing module processes the original documents and Wikipedia articles 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 [4]. 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.

At first we intend to use Wikipedia online. But the retrieval rate is closely depending on the performance of the Internet. For the sake of reducing the online retrieval time, we download the compressed XML format files from the Wikipedia website, extract text content of each article and index them to the local disk for the later retrieval.

There are several types of files contain different content on Wikipedia website. Now we only take the current pages file into consideration. This type of files contains current versions of article content. According to different time duration, different files are preserved on the website. In our experiments, we download the latest file. When the pages are revised, we can download it again and re-index them to keep the up-to-date information.

(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. We segment the question into words based on the same dictionary in the index processing module.

(3) Query expansion

We use the initial query expression to find the related article in Wikipedia. The query expansion module extract terms from the relevant paragraph, and then the new terms will be added to the initial query expansion to construct a new one for the retrieval phase on the given document.

(4) Retrieval

Our system employs the Vector Space Model (VSM) to determine the relevance between the given documents and the user's query [5].

There are two retrieval phases in our system. The related Wikipedia article can be gained by the initial retrieval phase using the initial query expression. After query expansion introduced in the section 3, the second retrieval results will be obtained using the new expanded query expression.

3. QUERY EXPANSION

For each question, the related query term will be submitted to the system to get a related Wikipedia article from the Wikipedia articles index. And choose the relevant paragraph to make query expansion.

3.1. Question classification

The question type are include DEFINITION, BIOGRAPHY, EVENT, RELATIONSHIP, WHY, PERSON, ORGANIZATION, LOCATION and DATE in NTCIR-8. We classify different

questions to different types. For example, the question “《千里走单骑》和张艺谋是什么关系？” belongs to de type of RELATIONSHIP and“高仓健是谁？”is a question belongs to BIOGRAPHY type.

3.2. Article Retrieval

For each of the different types of question, we use a different template to extract key terms. And then take retrieval in the Wikipedia by the name entities to get the related article. Sometimes we can not find the final article in the beginning, at this time we should relocate the final article by the most relevant title. For instance, “什么是SARS病毒？” is a DEFINITION question. We don’t consider the word“什么是”in this question but take “SARS病毒” as a key term. And when we choose the term “SARS病毒” to retrieve in the Wikipedia index, we find the term does not exist in the set of Wikipedia titles, so we have to choose the most relevant title of the question. In this case, we select “严重急性呼吸系统综合症” in the end, and locate the final article in wikipedia.

Some questions may have more than one name entities, we use different name entity to retrieval in Wikipedia to locate different articles. And then, choose relevant paragraph from every article to make query expansion.

3.3. Paragraph Location

After locate the final article, we need to find the most relevant paragraph to answer the question because the article we find is always too long and most of the content may be irrelevant to the question.

The paragraph we choose according to the type of the question. When the type is DEFINITION, BIOGRAPHY, PERSON or ORGANIZATION, we select first paragraph for query expansion. For other types of the question, we select the paragraph according to the key terms. For example, the question“第76届奥斯卡最佳男主角是谁”. Firstly we use the name entity “第76届奥斯卡” to locate the article “第76届奥斯卡金像奖”. Secondly, we select the paragraph “最佳男主角” according to the key term “最佳男主角”.

Some questions, especially the type “WHY”, may not locate the paragraph to answer the question. At this time, we should use Baidu or Google to search the related document, and then locate the relevant paragraph like in Wikipedia article.

3.4. 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. The Co-Occurrence based query expansion approach (CO) and Metric Correlation based query expansion approach (MC). The former approach is derived from the same idea as LCA (Local context analysis) [6]. In the latter method we extract words for expansion based on the idea of metric correlation [7].

For example, the initial query terms of the question “李永波和中国羽毛球队是什么关系？” are “李永波”，“中国”，“羽毛球”，“球队”. According to the key term “李永波”,we find the Wikipedia article “李永波”. And finally we locate the first paragraph for query expansion. After the terms for expansion are extracted, they are added to the initial query. In the end, the query terms increased to “李永波”，“中国”，“羽毛球”，“球队”，“1962年”，“著名”，“羽毛球运动”，“球运”，“运动员”，“教练员”，“辽宁”，“宁人”，“现为”，“国家队”，“总教练”，“国家”，“乒羽”，“中心”，“副主任”. Obviously, different expanded terms have different significance, so different weight should be assigned to them. The way we compute the weight of each term is the same as our work in NTCIR-7. The equation is defined as follows.

$$w(q \mid Q_{new})$$

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

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.

4. EXPERIMENTS

We submitted three formal run files to NTCIR-8 and the official evaluation results (AFTER bug fix) of performance are listed in Table 1.

Table 1. Formal run experiment official results (AFTER bug fix)

Run	Analysis File	Mean AP	Mean Q	Mean nDCG
WUST-CS-CS-01-T	No	0.2694	0.293	0.4881
WUST-EN-CS-01-T	No	0.1037	0.1206	0.2815
WUST-EN-CS-02-T	WHUQA-EN-CS-03-T.xml	0.1435	0.1564	0.292

In Table 1, “Run” indicates the name of the run file. In the name “EN-CS” indicates topic language is English and document language is Simplified Chinese. The suffix “T” indicates the question title. The “Analysis File” means the question analysis files offered by other participators.

Comparing with other groups, our system does not achieve a good official result. We think the reason is that the approach to extract key terms from the original question and the indexing progress in the program may have some problems.

As our system segment the question into words based on the same dictionary in the index processing module. If the key term does exist in the dictionary, our system may not retrieve the related document. For example, the extract key terms from the question “请描述万景峰号事件？” are “万”,“景峰号”,“事件”. But in fact, the right key terms are “万景峰”,“号”,“事件”.

In our “EN-CS” work, we extract English key terms and then translate them into Chinese by Google translation. So the quality of the translation determines the performance of our EN-CS result. For example, the key terms from the question “Who is the best actor in the 76th Oscar's?” are “最佳”, “演员”, “76” and “奥斯卡”, almost the same as the Chinese question. But the key terms from the question “How is Lin Chi-Ling after she crushed?” are “林”, “智”,“玲”,“后”,“她”,“粉碎”. It is very different from the Chinese question.

5. CONCLUSIONS

In order to solve the problems of inappropriate key terms exacted from the initial question and term mismatch, we apply query expansion technique to get more useful key terms for the query based on related Wikipedia article content. 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 when the question types are DEFINITION, BIOGRAPHY, PERSON, ORGANIZATION, LOCATION or DATE, we can find the relevant paragraph easily. But for other types of the question, we may have to use some other resource like Baidu and Google. The Wikipedia is a document sets of description type, so it performs better for the explanation questions.

6. REFERENCES

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

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

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

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

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

[6] 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.

[7] Ricardo B.-Y., Berthier R.-N., et al. Modern Information Retrieval. China Machine Press[C] 2006, 89.