Document Re-ranking via Wikipedia Articles for Definition/Biography Type Questions[*]

Maofu Liu[a], Fang Fang[a], and Donghong Ji[b]

[a]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, P.R.China
liumaofu@wust.edu.cn, fangfang0402@126.com
[b]School of Computer, Wuhan University, Wuhan 430079, P.R.China
donghong_ji2000@yahoo.com.cn

Abstract. In this paper, we propose a document re-ranking approach based on the Wikipedia articles related to the specific questions to re-order the initial retrieved documents to improve the precision of top retrieved documents in Chinese information retrieval for question answering (IR4QA) system where the questions are definition or biography type. On one hand, we compute the similarity between each document in the initial retrieved results and the related Wikipedia article. On the other hand, we do clustering analysis for the documents based on the K-Means clustering method and compute the similarity between each centroid of the clusters and the Wikipedia article. Then we integrate the two kinds of similarity with the initial ranking score as the last similarity value and re-rank the documents in descending order with this measure. Experiment results demonstrate that this approach can improve the precision of the top relevant documents effectively.

Keywords: document re-ranking, Chinese IR4QA, Wikipedia, clustering analysis

1 Introduction

It is reported that most of the information system users are accustomed to browse the top returned search results only (iProspect, 2004), so they hope the top ranking documents are highly relevant. In order to meet the information need, it is necessary and significant to improve the precision of top retrieved documents. Currently, document re-ranking has become one of the main streams to improve the precision of top retrieved documents. After document re-ranking, it is expected that more relevant documents appear in the higher rankings.

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 existence of query terms, but semantics of query terms in documents. If the IR system just simply checks the existence of query terms in documents without considering the context of documents, it often causes term mismatch and declines the performance greatly (Salton and McGill, 1983). So the important problem in automatic document re-ranking is the relevance measure of document and query. The strategies, such as query expansion, latent semantic indexing (LSI) and mutual information, have been proposed to solve this problem. And it has been proved that these approaches can improve the performance of the retrieval system effectively.

[*] The work in this paper was supported by Research Grant from Hubei Provincial Department of Education (No. 500064).

Geetha and Kannan (2007) put forward another approach to solve this problem. When the initial results are returned, the user can choose a document of interest as the seed document and initiate the re-ranking algorithm by which documents are re-ranked based on its similarity distance from the seed document. This algorithm helps users to re-rank documents based on seed document as a query. But it needs the interaction of users.

In this paper, we make use of the related Wikipedia articles to calculate the Wiki-Document and Wiki-Cluster similarities to adjust the ranking score for document re-ranking to solve the problem mentioned above.

As we know, Wikipedia[1] is a multilingual, web-based, free-content encyclopedia project, and each of its article provides information to explain the term of the article title. We can make use of the related Wikipedia article as the "seed" document of a query, without choosing the seed documents manually, mentioned by Geetha and Kannan. We calculate the similarity distance between the seed document and each document in the initial retrieved results, named **Wiki-Document similarity**.

In the IR4QA task of 7th NTCIR, we look on the question as the query. The whole Wikipedia article states like the definition of its title, so we just take the specific type of questions, the definition/biography type[2] into consideration. As the name implies, the definition/biography type questions are the ones about definitions of terms or biographies of persons, such as "What is the Nobel Prize?", "Who is Osama bin Laden?".

Many researchers have made efforts on how to apply clustering to get better retrieved results. The document clustering based approach is now a typical one of document re-ordering, which is based on the assumption that cluster related documents should be more similar to each other and the content similar documents may appear in the same category with greater possibility (van Rijsbergen, 1979). Document clustering approach assigns documents to automatically created clusters, based on the degree of association between documents and clusters. But this is inappropriate to calculate the similarity of the query and the document since query consists of only a few terms for obtaining statistically meaningful frequency-vector and clusters.

In order to apply the document clustering analysis for document re-ranking, we also make use of the related Wikipedia article as a question. When the initial ranking documents are divided into clusters, we can calculate the similarity between the question-related Wikipedia article and the document cluster, instead of the question and document cluster, named **Wiki-Cluster similarity**, to regulate the ranking score.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 introduces the proposed document re-ranking approach using Wikipedia article. Section 4 then presents experiment results. Finally, Section 5 concludes the paper.

2 Related Works

Many methods have been put forward to re-rank the initial retrieved documents. Lee *et al.* (2001) proposed a clustering analysis method which uses static and dynamic cluster views to re-rank the documents. It is reported their method achieves significant improvements on Korean corpus. Shi (2005) applied a boosting algorithm that captured natural language substructures embedded in texts to re-rank the retrieved documents. Experiment results show that the boosting algorithm worked well in cases where a conventional IR system performs poorly, but the re-ranking approach was not robust enough when applied to broad coverage task typically associated with IR. Kemps (2004) proposed a method to re-order retrieved documents by making use of manually assigned controlled vocabularies in documents. And it is reported that this re-ranking strategy significantly improves retrieved effectiveness on their experiments on German GIRT and French Amaryllis collections. Balinski and Danilowicz (2005) put forward a

[1] http://www.wikipedia.org

[2] http://aclia.lti.cs.cmu.edu/wiki/TaskDefinition#Format

document re-ranking method that uses the distances between documents for modifying initial relevance weights. Yang *et al.* (2004, 2005) used query terms which occur in both query and top N (N<=30) retrieved documents to re-rank documents.

Many research efforts have been made on how to apply clustering to get better retrieved results. Lee *et al.* (2001) proposed a model of information retrieval system that is based on a document re-ranking method using document clusters mentioned above. Anick and Vaithyanathan (1997) exploited the synergy between document clustering and phrasal analysis for the purpose of automatically constructing a context-based retrieval system. In their system, a context consists of two components, cluster of logical related articles (its extension) and a small set of salient concepts, represented by words and phrases and organized by the cluster's key terms (its intension). The Scatter/Gather system (Hearst and Pedersen, 1996) was a cluster-based document to browsing method, as an alternative to ranked titles for the organization and viewing of retrieved results. All of their experiments show that the clustering methods can enhance the performance of the IR systems.

3 Document Re-ranking via Wikipedia Articles

3.1 System Architecture

For each question, the related question/query term will be submitted to the system to get a relevant Wikipedia article from the Wikipedia articles index, which will be explained in detail in the following section for each document. Then the Wikipedia article and the initial retrieved documents are input into the document re-ranking module. This module will compute the Document-Wiki similarity and the Wiki-Cluster similarities. When these two similarities are gained, they are combined with the initial score to generate the last similarity score. Then, documents are descending ordered by the last score and the final ranking documents are returned to users.

The system architecture is illustrated in Figure 1.

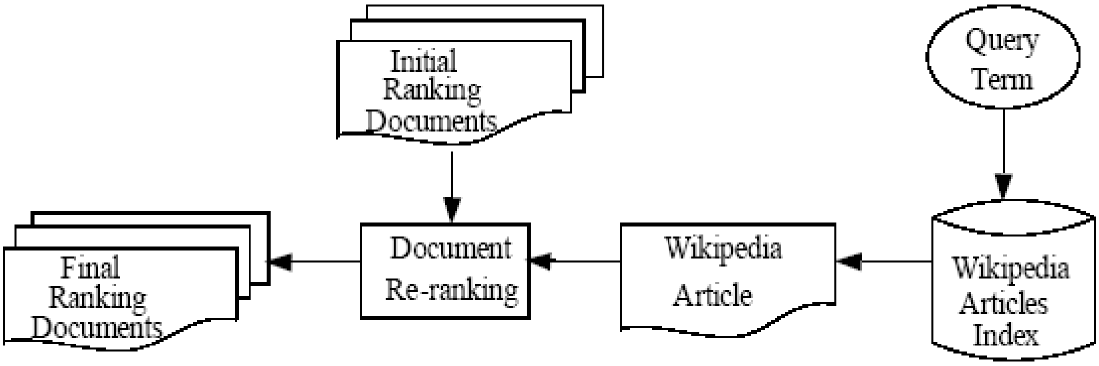

Figure 1: System architecture

3.2 Wikipedia Resource Processing

For a question of definition/biography type, we want to get the related pages of the Wikipedia article which the title is the key term of the question. For example, if the question is "What is the Nobel Prize?", we would like to obtain the text content of the page which the title is "Nobel Prize". It is feasible that searching online by submit the key terms in the question to the Wikipedia search portal while the retrieval system is running. But the retrieval rate is closely depending on the performance of the network. For the sake of reducing the retrieval time, we download the compressed xml format files on the related Wikipedia site, extract text content of each article and index them to the local disk for the later retrieval. Due to the Wikipedia resources are on the local disks, the retrieval rate will be greatly improved than searching the information online.

There are several types of files contain different content on Wikipedia site. 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 site. 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.

3.3 Wiki-Document Similarity

For each document in the initial retrieved results, we will calculate the similarity with the related Wikipedia article, the Wiki-Document Similarity. There are several models to compute the similarity. We choose the well-known Vector Space Model (VSM). After the text contents of each document including the Wikipedia articles are segmented by a Chinese segmentor ICTCLAS[3], documents are represented by a vector in vector space where each dimension of vector is a word. We use the tf*idf values as the weighting scheme for the words. The same document representation and the weighting strategy for words will also be used for calculating the following Wiki-Cluster Similarity. So the similarity of document *d* in the initial retrieved results with the Wikipedia article can be computed by the following equation.

$$Sim(w,d) = \cos(\vec{w},\vec{d}) = \frac{\vec{w}\cdot\vec{d}}{\left|\vec{w}\right|\cdot\left|\vec{d}\right|} \tag{1}$$

Where $Sim(w,d)$ expresses the Wiki-Document similarity, and $\vec{d}$ and $\vec{w}$ denotes the vector of document d and the Wikipedia article respectively.

3.4 Wiki-Cluster Similarity

The clustering analysis hopes to divide the similar documents into the same cluster, and then uses the correlation of the question and the cluster information to regulate the sorting score. It is likely to that the document *A* gains high relevant score when it contains the key terms of the question while document *B* gets low score for it doesn't contain these key terms although its content is related to the question. When *A* and *B* have similar contents as they contain other same terms, they would be grouped into the same cluster. So when the cluster has high similarity with the question, the document *B* can get higher similarity score to enhance the ranking position based on the correlation of the cluster.

Several document clustering approaches to cluster document set have been put forward. K-Means and hierarchical clustering are the two typical ones. Here we make use of the K-Means cluster method (Ren *et al.,* 2006) to do the clustering analysis and assign the cluster similarity score for each document. The process is listed as follows.

1) Create the vectors of the Wikipedia article and each of the documents in the initial results.
2) Utilize the K-means algorithm to do clustering analysis for the documents in results and calculate the centroid of each cluster.
3) Compute the similarity of the Wikipedia article vector and each of the cluster centroid with the cosine coefficient measure.
4) Assign the similarity value $Sim(w,c_i)$ to each of the documents.

$$Sim(w,c) = \cos(\vec{w},\vec{c_i}) = \frac{\vec{w}\cdot\vec{c_i}}{\left|\vec{w}\right|\times\left|\vec{c_i}\right|} \tag{2}$$

[3] http://ictclas.org/

Where $Sim(w, c_i)$ denotes the Wiki-Cluster Similarity and $\vec{c_i}$ means the centroid of the i^{th} cluster.

When do the K-Means clustering, we set the number of the cluster *K* according to the number of the initial retrieval result documents. Following is the formula to initialize this parameter.

$$K = N / n \tag{3}$$

Where *N* is the total number of the result documents for a question, *n* is the experimental determined parameter which is taken as the average documents number of all the clusters. In our experiments, *n* is set to 5, and *N* is 50.

3.5 Document Re-ranking

We combine the similarity $Sim(q, d)$ between each document and the question in the initial results with the Wiki-Document similarity and the Wiki-Cluster similarity. The formula is displayed as follows.

$$\begin{aligned} &Sim(d) = \alpha \cdot Sim(q, d) + \beta \cdot Sim(w, d) + \gamma \cdot Sim(w, c_i) \\ &\alpha + \beta + \gamma = 1 \end{aligned} \tag{4}$$

Where $Sim(d)$ implies the final similarity of document d . α , β and γ are parameters to adjust the different importance to each of the similarities and assigned to 0.6, 0.3 and 0.1 respectively in our following experiments,.

When the last similarity is calculated, the documents are re-ranked by descending order of this measure and the final results are returned to users.

4 Experiment Results

We evaluate the proposed document re-ranking approach based on related Wikipedia articles on Simplified Chinese document collections for the 7th NTCIR ACLIA/IR4QA task which are from Xinhua Chinese and Lianhe Zaobao between the year 1998 and 2001. The total number of the documents is 539,062.

In question set of the 7th NTCIR ACLIA/IR4QA task, there are four types of question and totally 98 questions. There are two different fields to describe one question, title and narrative, and we only choose the title field to represent the question in the experiments. In this paper, we select 36 definition/biography questions, which can obtain related articles on the Wikipedia site. The questions and the Wikipedia articles are all in Simplified Chinese.

The initial retrieved results are generated by the approach put forward by Liu *et al.* (2008) and we select the top 100 documents in the results for the further re-ranking.

We mainly use the P@N evaluation criterion to analyze our experiments. Two kinds of relevant measures, relax relevant and rigid relevant, are concerned. The rigid relevant means the document is highly relevant or relevant with a question, and relax relevant denotes the document is highly relevant or relevant partially with a question. Table 1 lists the baseline results, representing the initial results, and the re-ranking results, obtained by using our re-ranking approach.

Table 1: P@N in relax and rigid measure

P@N	Baseline	Re-ranking	Improvement (%)
P@10(relax)	0.339	0.383	13.0
P@10(rigid)	0.136	0.194	42.6
P@20(relax)	0.343	0.364	6.12
P@20(rigid)	0.111	0.133	19.8
P@30(relax)	0.311	0.327	5.14
P@30(rigid)	0.103	0.116	12.6

From Table 1, we can see that our approach can enhance P@10 by 5.1% from 0.339 to 0.383 in relax relevant measure while improve 42.6% from 0.136 to 0.194 in rigid relevant measure. Other groups of data such as P@20 and P@30 are also displayed in the table with the measure of relax and rigid respectively. From the groups of the relax measure and the rigid measure, we can see that the improvement of rigid measure are better than the relax measure when evaluating with different P@N.

Figure 2 below shows the improvements in measure of rigid P@10 for each question. The x-axis denotes the No. of each question.

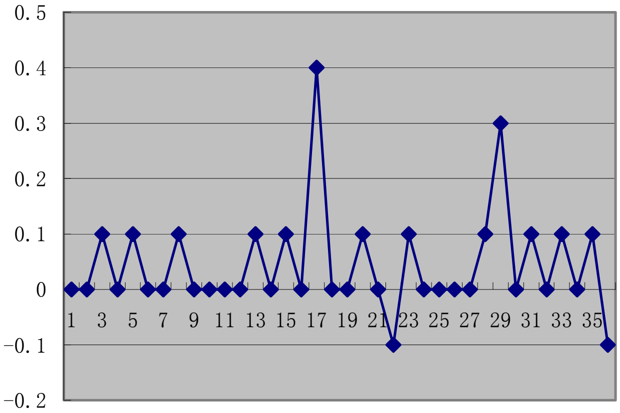

Figure 2: Increments in rigid P@10 measure

From Figure 2, we can see that most of the questions are better in rigid P@N measure after re-ranking while a few of them are worse, such as Question 22 and 36. Look into the two Wikipedia articles related to these two questions, we find that there are some redundant contents, such as the development history of the term, which takes up long length of the whole document. The terms in this content may be irrelevant with the question so it declines the precision. So, we can conclude that the content of the Wikipedia articles have great impact on the re-ranking results.

5 Conclusions

In this paper, we put forward a re-ranking approach via question related Wikipedia article for the definition/biography type question. We make use of the related Wikipedia articles to calculate the Wiki-Document and Wiki-Cluster similarities to adjust the ranking score. The experiments show that our approach can improve the top precision for the IR4QA system. The improvement of P@N in rigid measure is higher than the relax measure where the rigid from 12.6% to 42.6% and the relax measure from 5.14% to 13.0% with different P@N evaluations.

The re-ranking approach in this paper is applied to specific type of questions for the convenient acquisition of the whole article content from the Wikipedia. But the content of some articles does not well state the related terms as the seed document. If we choose these articles, it may decline the retrieval precision. There are other free encyclopedia resources such as the Baidu encyclopedia. We may integrate some of these encyclopedias to generate the model answers to be used as a seed document.

References

Geetha, A. and A. Kannan. 2007. Enhancement of Search Results Using Dynamic Document Seed Reranking Algorithm. *Journal of Computer Science,* 3(6):436-440.

Anick, P.G. and S. Vaithyanathan. 1997. Exploiting clustering and phrases for content-based information retrieval. In *proceedings of 20th ACM SIGIR International Conference on Research and Development in Information Retrieval*. pp.314-323.

Balinski, J. and C. Danilowicz. 2005. Re-ranking method based on inter-document distance. *Information Proceeding and Management,* 41,759*-775.*

Hearst, M.A. and J.O. Pedersen. 1996. Re-examining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of 19th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 76–84.

iProspect. 2004. iProspect's Search Engine User Attitudes Survey Results [DB/OL]. http://www.iprospect.com/.

Kamps, J. 2004. Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary. In *Proceedings of the 21th European Conference on Information Retrieval.*

Lee, Kyung-Soon, Young-Chan Park and Key-Sun Choi. 2001. Re-ranking model based on document clusters. *Information Processing and Management,* 37*,* 1-14.

Liu, Maofu, Fang Fang, Qing Hu and Jianxun Chen. 2008. Question Analysis and Query Expansion in CS-CS IR4QA. In *Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan.*

Ren, Jiangtao, Jinghao Sun, Xiaoxiao Shi *et al.* 2006. An Improved K-means Algorithm for Text Clustering. *Computer Applications,* Vol 26.

Salton, G. and M. McGill. 1983. *An introduction to modern information retrieval [M]*. New-York: McGraw-Hill.

Shi, Zhongmi, Baohua Gu, Fred Popowich and Anoop Sarkar. 2005. Synonym-based Query Expansion and Boosting-based Re-ranking: A Two-phase Approach for Genomic Information Retrieval. In *the Proceedings of TREC2005.*

van Rijsbergen, C.G. 1979. *Information Retrieval*. Butterworths, London, second edition.

Yang, L.P., D.H. Ji and L. Tang. 2004. Document Re-ranking Based on Automatically Acquired Key Terms in Chinese Information Retrieval. In *Proceedings of 20th International Conference on Computational Linguistics (COLING).*

Yang, L.P., D.H. Ji, G.D. Zhou and Y. Nie. 2005. Improving Retrieval Effectiveness by Using Key Terms in Top Retrieved Documents. In *Proceedings of 27th European Conference on Information Retrieval.*