

Chinese Information Retrieval Based on Related Term Group

Tingting HE¹ Guozhong QU¹ Xinhui TU¹ Donghong JI²

¹Department of Computer Science, Huazhong Normal University Wuhan, 430079
hett@mail.ccnu.edu.cn qu_g_z@mails.ccnu.edu.cn tuxinhui@mails.ccnu.edu.cn

²Institute for Infocomm Research Heng Mui Keng Terrace, 21 Singapore 119613
dhji@i2r.a-star.edu.sg

Abstract

This paper describes our work at the fifth NTCIR workshop on the subtasks of monolingual information retrieval (IR). Query expansions using automatically acquired related term groups were explored. Unlike traditional query expansion methods, the related term groups extracted from web-based corpuses and the related terms extracted from document set are used in combination to improve the effectiveness of query expansion in our method. Experiments show that our method achieves an average 13.1% improvement compare to the traditional relevance feedback technique.

Keywords: related term group, information retrieval.

1 Introduction

One major problem in information retrieval is term mismatch between queries and documents [1]. The problem of term mismatch in information retrieval occurs because people often use different terms to describe concepts in their queries than authors use to describe the same concepts in their documents [2]. 秘笈/要诀/绝招/法宝 are words used in the HTML documents related to 秘诀, which may vary from one document to another. If a user uses a query with word 法宝, he/she cannot retrieve relevant documents due to the term mismatch. Query expansion has been suggested as a technique for dealing with this problem by expanding queries using semantically similar and syntactically related words to those in the query to increase the chances of matching terms in relevant documents.

Currently, many information retrieval researchers use pseudo relevance feedback, which is relevance feedback without user intervention [3][4]. Firstly, several documents are retrieved as a result of the initial retrieval. Assuming that the top-n retrieved documents are relevant, the system uses the terms contained in those documents as expansion terms and retrieves again. This technique works well when the system has a fairly good performance and the top-n

retrieved documents are really relevant. If this assumption does not hold, however, the performance of the system can degrade as the expansion terms are taken from irrelevant documents.

Query expansion using thesauri is still worth investigating. There are two types of thesauri, that is, handcrafted thesauri [5] and automatically constructed thesauri [6][7][8]. WordNet [9] is an example of a handcrafted thesaurus, which is available in machine-readable form. Many researchers have tried to use relations defined in WordNet for query expansion. Unfortunately, the results have not been as good as expected [10].

In this paper, we propose a novel method to improve the performance of Chinese information retrieval systems by expanding queries using automatically acquired related term groups. Unlike traditional query expansion methods, the related term groups extracted from web-corpora and the related terms extracted from document set are used in combination to improve the effectiveness of query expansion in our method.

The experiments described in this paper were not finished at the result submission. The results we have submitted are based on bi-gram index and pseudo relevance feedback. Now we have finished the experiments, so in this paper we primarily discuss our new method.

The rest of this paper is organized as following. In section 2, we describe indexing method and automatic related term group acquisition. In section 3, we demonstrate the process of query expansion. In section 4, we evaluate the performance of our method on NTCIR test collections and give out some result analysis. In section 5, we present the conclusion and some future work.

2 Preprocessing

2.1 Indexing

For Chinese information retrieval task, bi-gram and word both are the most effective indexing units

[11][12][13]. However, they are not ideal units for query expansion. We use automatically extracted terms as indexing units in our work in order to improve effectiveness of query expansion.

To acquire Chinese terms, we use a clustering-based term extraction method [14][15]. We first roughly cluster the whole document set r into K ($K < 2000$) document clusters by using K-Means, then we regard each document cluster as a large document and apply term extraction algorithm on each document cluster and respectively get terms in each document cluster. We use a Pat-tree based method to automatically extract term from each document cluster [16][17]. All these terms from different document clusters form the whole terms list. We regard a term whose length is less than 4 Chinese Characters as a short term, and a term whose length is equal or greater than 4 Chinese Characters as a long term. Following is some examples of short terms and long terms.

(1) Short Terms

- 丑闻(Chou3 Wen2)
- 地震(Di4 Zhen4)
- 政策(Zheng4 Ce4)

(2) Long Terms

- 奈米技术(Nai4 Mi3 Ji4 Shu4)
- 粮食危机(Liang2 Shi2 Wei2 Ji1)
- 货币政策(Huo4 Bi4 Zheng4 Ce4)

2.2 Related Term Group Acquisition

In traditional corpus-based query expansion, the term-term similarity matrix is often used to improve retrieval effectiveness. Various similarity measures have been suggested, such as cosine similarity, mutual information etc. In our approach, we get the related terms by using the term-term similarity matrix, and furthermore, we cluster all related terms of one short term into several groups. The terms in the same group is only related to one topic. These related term groups will be used to do query expansion.

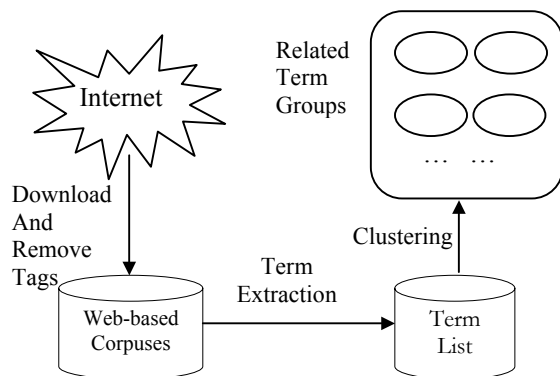


Figure 1. Process of related term group Acquisition.

Figure 1 illustrates the outline of related term group acquisition. At first, we download about 9.8GB web pages from 31 important news web sites. The web pages are converted into text files in the same Chinese code by removing the HTML tags. According to different channel, which the web pages belong to, we construct 17 corpuses belong to different topic. Table 1 gives the detail description of the corpuses.

For each corpus, we can get a term collection through term exaction. Then, a term-term mutual information matrix is constructed for each corpus. In order to get more accurate relevance information, we use passage, not entire document, as unit to calculate mutual information. If the normalized mutual information of two terms is greater than a constant z , we consider that the two terms are related. According to this criterion, we can construct a related term collection for each term.

We find that the terms in a related term collection usually belong to several different topics. Using all terms in the collection to do query expansion probably deteriorate the effectiveness of query expansion. Therefore, according to the mutual information criterion we mentioned above, we cluster the terms in a related term collection into several groups, the terms in the same group belong to same or similar topic and different group belong to different topics.

Clustering is a time-consuming process. It will consume much time to generate related term group for all terms. In addition, long terms usually belong to some special domain; they have very few related terms. Considering the facts above, we only generate related terms group for the short terms occur in the queries.

Topic	Size(MB)
Politics	1,141
Military affairs	850
Sports	615
Finance and economics	718
Education	646
Health	623
Science and technology	727
Culture	586
Entertainment	678
Tour	373
Game	436
Automobile	395
House	437
Mobile telephone	384
Business	426
Live	327
Movie	436
Total	9,798

Table 1. Description of the corpuses.

Mutual information is a significant indicator of the relevance of two terms. To cluster highly interrelated terms into the same group, the reciprocal of mutual information is adopted as the measure of distance between two terms.

Following is the procedure to construct related term groups for one short term.

Input: term-term mutual information matrix M , term T

Output: related term groups of term T

Step 1: For each term T_1 in M , we add T_1 to the related term collection of T , if the mutual information between term T and term T_1 is greater than a constant.

Step 2: Calculate the distance of each term pair in the related term collection. We regard the reciprocal of the mutual information of two terms as the distance between them.

Step 3: We cluster all terms in the collection into several groups by using group-average agglomerative clustering algorithm.

3 Retrieval processing

3.1 Retrieval

Figure 2 demonstrates the retrieval process of our Chinese information retrieval system. Firstly, we automatically extract terms from test document set and use them to build indexes. Secondly, we build a few of web-based corpuses belonging to different topics and automatically extract related term groups from them. Thirdly, we make use of terms in query and document to do initial search to get initial ranking document. Fourthly, we construct related term groups for each short term in query. Fifthly, we make use of related term groups of each short term in query together with related terms of the terms in query and top N documents in initial ranking documents to do query expansion to form a new query. Finally, we use the new query to search again to get final ranking documents.

3.2 Query expansion

Our system makes use of the information of top 30 initial ranking documents together with pre-built related term groups, short terms and long terms in query and their relevant terms (co-occurred terms) in document set to do query expansion.

Following is the procedure to expand a query q :

Step 1: Each short term t in query q may occur in several different corpuses we constructed in section 2.2. In order to choose which corpus to generate related term groups, query weights are calculated for each corpus based on term frequency and inverse document frequency. The weighting method is as follows:

$$weight(i, j) = (1 + \log(tf_{ij})) \log(N/df_i) \quad (1)$$

$$qweight_j = \sum_{i=1}^n weight(i, j) \quad (2)$$

Where tf_{ij} is the occurrence frequency of term t_i in corpus C_j . N is the total number of documents in all corpuses, and df_i is the number of documents, which contain term t_i .

For a query, corpus C_j will be chosen, if $qweight_j$ is the greatest. For each term t in query q , $G(t)$ is one of the related term groups of t in corpus C_j , all terms in $G(t)$ will be added into q if at least two terms in $G(t)$ occur in the top 30 ranking documents with 0.5 as weight.

Step 2: For each term t in query q , $R(t)$ is relevant term of t , $R(t)$ will be added into q with 0.5 as its weight if $R(t)$ occurs at least in two document among the top 30 ranking documents. We acquire relevant terms of term t by their co-occurrence in documents and their mutual information.

Step 3: All terms in q are added to new query with frequency in q as its weight.

The original query plus new terms acquired by query expansion form a new query. This new query is used to search again to get final search result.

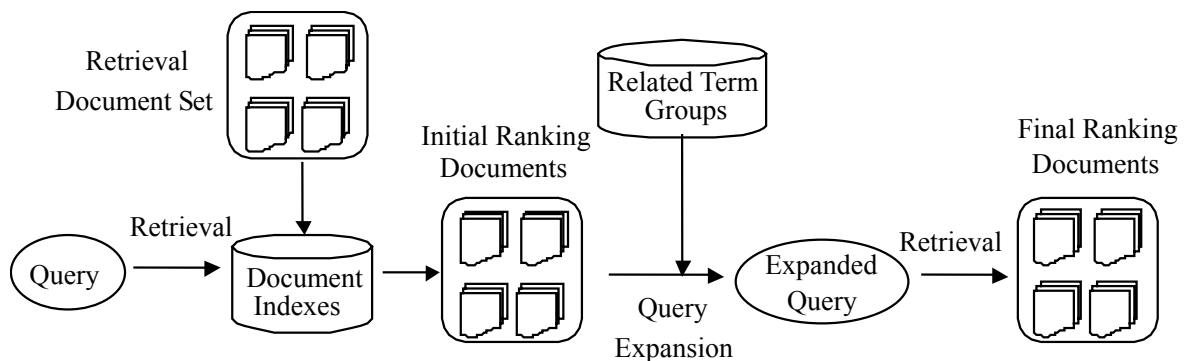


Figure 2. Process of retrieval.

