Chinese Language IR based on Term Extraction

Ji Donghong, Yang Lingpeng, Nie Yu
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613
{dhji, lpyang, ynie}@lit.a-star.edu.sg

Abstract

In this paper, we'll describe the core technology and modules we use in LIT (formerly KRDL)'s Chinese Language Information Retrieval System. The system mainly includes automatic term extraction from Chinese documents, query analysis based on the terms and finally measurement of the association between queries and documents. Compared with other methods, we try to use automatically acquired terms as well as their related terms as features to retrieve documents, so we don't need any word segmentation procedures as prerequisite. The terms are more significant than words in representing the queries.

Keywords: *Term Extraction, Relevant Terms, Query Expansion, Query Analysis, Information Retrieval*

1 System Description

Generally there are two ways to tackle the Chinese characters in Chinese information retrieval. One way is to use Chinese characters directly as features of documents, but such a *bag* of Chinese characters, not considering the ordering of the characters, may reveal too little information about the semantic content of documents because that the number of the Chinese characters are very limited compared with the number of the words in English, thus the ordering of the Chinese characters is very important in conveying information. The other way is to use words as features of documents, which needs some word segmentation as a prerequisite, however, word segmentation still presents a major difficulty in Chinese information processing [8]. In our system design, we try to use automatically acquired terms [3, 6] as features of documents. In detail, we first extract key terms as well as their related terms from documents, then conduct query analysis based on these terms, finally use query terms as features to retrieve documents.

Our Chinese Information Retrieval System is a three-step system (see fig. 1). The first step is to extract key terns and their relevant terms from the document collection, the second phase is to determine key terms from each query and the last phase is to retrieve relevant documents based on the key terms found in queries and documents.

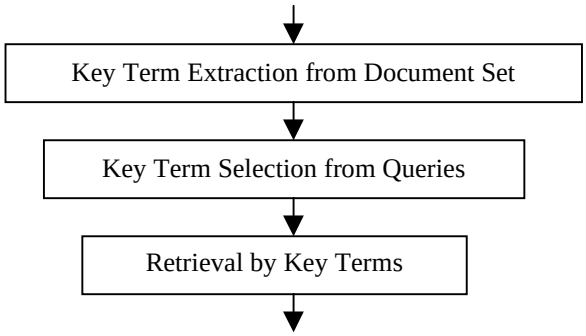

Figure 1. Diagram of System

2 Key Term Extraction from Document Set

We extract key terms from the document set. A key term in a document is a Chinese character string that occurs statistically importantly in a document [1, 2, 3, 4]. They always have much contribution to the semantic content of documents. A key term is generally different from a Chinese word derived via traditional word segmentation procedures. In fact, we don't use any segmentation algorithm here.

At the same time, we acquire relevant terms for each key term. Relevant terms refer to those terms that often co-occur with one key term in documents. Intuitively, they usually carry relevant meanings of key terms, so they can be used for query term expansion.

Regarding term extraction, there have been many strategies now ranging from rule-based methods to statistical methods or their combination [4]. Since both word segmentation and part-of-speech definition are still very controversial and difficult tasks in Chinese information processing currently, we adopt a kind of pure statistical strategy, i.e., "seeding and extension" strategy: first determining some *seed* characters for possible terms in the documents [7], and then extending to

the contexts of the seeding characters to find appropriate terms. Following is the general procedure of extracting terms from a given document in the document set:

- To determine character seeds from document by comparing the frequency of the character in the documents against that of the same character in the reference corpus. Ideally the reference corpus should be very large and balanced in order to determine better seeding words, however, such a corpus is not available to us, so we just use the document collection as the reference corpus.

- For each selected character seed, to check its every occurrence in document to see if it can be expanded to the left or to the right to form a new term making use of mutual information. If a new term can be formed, we put the new term into term candidate set (T).

- For each term candidate in set T, we do the similar processing as we do for character seeds to identify more new term candidates until no new term candidates can be formed.

- To scan set T to determine the key terms based on the strategy of maximal length preference.

In general, a key term is a Chinese character string that contains at least one seed character and among which the characters meet the mutual information criteria.

3 Key Term Selection from Queries

After extracting key terms and their relevant terms from documents, we can scan each query to check which part of the query is a key term or is the sub-string of a key term. We call the above key terms or the sub-string of key terms as query key terms. All these query key terms and their relevant terms are used to construct a query vector to formally represent the query, the vector is called query key term vector.

For each query key term in a query vector, it'll be assigned a weight depending on its length, position in the query sentence and its relative frequency in the document vs. in the whole document collection. Generally speaking, the longer the term is, the higher its weight, and the higher the relative frequency, the higher its weight. Regarding the position factor, we now only consider some special prepositional structures, e.g., structures including 在 (in), 的(of), 中(among), 對(to), 之(of) etc.

Specially, we do some statistical analysis of all retrieval topics to determine which query key terms in retrieval topics may be just a description term. A description term will be assigned lower weight. For example, those terms that occur in more than four retrieval topics and mostly occur in the rear or in the front of retrieval topics are mostly possible to be description terms.

In fact, this module can be seen as a special kind of word segmentation. However, the segmentation procedure here is only based on the terms we automatically acquired from the document collection, rather than a pre-defined word segmentation dictionary as usual. On the other hand, we expect that these terms may be more significant in representing the queries than those words acquired by traditional word segmentation algorithms, because that the terms are relatively frequent in the documents, thus carrying more semantic content of the documents, and that a term, often consisting of multi-words, generally represent more complete concepts than a word does.

4 Document Retrieval by Key Term Vectors

For each query, we scan all the documents in document set to check if it matches this query. In the process, each document is also represented by a vector based on extracted terms, their weight is defined by traditional tf/idf scheme [5]. Then we check the distance between the query and the document based on the cosine of the two vectors. If the distance between two vectors is close, we'll select the document as a possible query answer.

5 Discussion of Results and Future Work

Although terms are promising alternative of words or characters as features to retrieve Chinese documents, there is much room for improvement in the performance of our system. Although we want to try this new method, the experimental result is not as good as expected. After checking the results, we find that the main reasons for the retrieval errors are as the following:

1) Term extraction errors. We need a large corpus as a reference corpus in order to determine better seed characters, and thus improve the performance of term extraction, but such a corpus is not available to us. So, some really good key terms are not identified in the final result.

2) Query key word expansion. Some relevant documents cannot be retrieved unless you can get good relevant terms of the query terms. But some key relevant terms cannot be identified only based on co-occurrence

statistics. We need to use more fine-grained lexical resources, e.g., thesaurus or domain-specific knowledge base. But such fine-grained lexical resources are generally unavailable now.

3) The weight of the terms in query vector. Now the weight is dependent on the length, position and frequency of the terms in the corpus. But the formulation of these factors and their combination should be more fine-grained.

4) The distance threshold we adopt in our system for retrieving documents is too high, so for many queries, less than 100 documents are retrieved. That's why there is no improvement in our final result when comparing Pre100 and Pre1000 with Pre10.

In future, we still plan to adopt automatically acquired terms, rather than words, as features to retrieve documents. We need to consider the following problems in more detail.

1) To improve the performance of term extraction and relevant terms acquisition. In addition to the document collection, we can make use of internet as a dynamically updated resource to acquire relevant terms, and then combine them with those identified from the documents to form better and more complete relevant terms set.

2) To improve the formulation of the queries based on the acquired terms and relevant terms. Some terms in the query are very important, some are less important, and others may be just descriptions. Sometimes, some relevant terms may be very important. So, how to formally evaluate their contribution to the query is a key for correct document retrieval.

3) Since the queries are too short in general, so their formulation as vectors may be not very appropriate, we may consider to use some score function to evaluate the relevance between the documents and queries. However, we need to incorporate the effect of relevant terms as query expansion, whatever the cost function is.

References

[1] L. F. Chien, M. C. Chen, C. L. Chen, B. R. Bai. 1998. Internet-based text corpus classification and domain-specific keyterm extraction. *Proceedings of First Workshop on Computational Terminology*. pp: 71-75.

[2] K. Frantzi, S. Ananiadou, H. Mima. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *Journal of Digital Library*. 3: 115-130.

[3] Fung, P. 1998. Extracting key terms from Chinese and Japanese texts. *The International Journal on Computer Processing of Oriental Language. Special Issue on Information Retrieval on Oriental Langauages*, pp 99-121

[4] G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston.

[5] Manning, C. D. and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

[6] Patrick Pantel and Dekang Lin. 2001. A Statistical Corpus-Based Term Extractor. In: Stroulia, E. and Matwin, S. (Eds.) *AI 2001, Lecture Notes in Artificial Intelligence*. pp. 36-46. Springer-Verlag.

[7] H. Schutze. 1998. The hypertext concordance: a better back-of-the-book index. *Proceedings of First Workshop on Computational Terminology*. pp: 101-104.

[8] Sun, M., Shen, D. and Tsou B. K. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of COLING-ACL-98*. Pp 1265-1271. Montreal, Canada.