Exploration and Study of Chinese Thesaurus Automation Construction for Digital Libraries

Wen ZENG, Wen ZENG
Institute of Scientific and Technical Information of China, zengw@istic.ac.cn

Abstract

The paper aims to explore Chinese thesaurus automation construction based on the freely available digital library resources. The key methods and study results are presented in the paper. The study adopted the technology of natural language processing to analyze the linguistic characteristics of terms, and combined with statistical analysis to extract the terms from technical literatures. Our method consists of automatic extracting and filtering terms, judging and building relationship among terms. The study experiments run on the Java test platform. The study obtains the following conclusions: Finding the similarities and differences between the Chinese thesaurus standard and International thesaurus standard. The methods for the automatic extraction of terms and building relationship among them are presented. The study results show that the proposed methods can obtain better performance. The paper proposes a new thinking on thesaurus automation construction in the digital age. The presented method based on linguistics and statistics is a new attempt. According to the experiment results, this exploration and study is innovative and valuable. In addition, these thinking and methods give a good start for improving information services of the PRC’s National Science and Technology Digital Library.

Keywords: Digital libraries, Chinese thesaurus, Automation construction, Term extraction, Linguistics

1. Introduction

Digital libraries are important that preserve culture and technology information via development of networks. The emergence of a large number of information resources has accumulated a wealth of digital information. The traditional internet technology is only a connection to network resources and does not take into account the structural organization of network resources, resulting in a variety of knowledge and thousands of storage media disorderly, fragmented distribution. How to quickly find information from vast amounts of digital information resources that user needs is an unsolved problem. In order to solve this search problem, the core problem lies in finding an orderly method of organizing information resources. Thesaurus is a tool for information organization. Thesaurus construction is of an important and significance work for the service of digital library. It can achieve efficient organization and utilization of digital literatures resources. Because of facing the mass literature resources, traditional Chinese thesaurus has failed to meet the library personnel for literatures processing demand. Library personnel requires new thesaurus for indexing and retrieval, users also need to use practical thesaurus to improve literature retrieval efficiency and quality.

The People's Republic of China has made a great effort to promote the development of digital collections and exploiting. The quality and quantity of digital collections have improved dramatically after development and promotion of nearly ten years. National Science and Technology Digital Library (NSTL) was established in 2000. Today, it is becoming a national engineering technology of digital library. As its member unit, Institute of Scientific and Technical Information of China (ISTIC) provides technology support for its service and running. The Chinese thesaurus was edited and published in 1980, and its revision was published in 1996. So, Chinese thesaurus has fallen behind the times, it needs to be constructed again. How to construct Chinese thesaurus effectually is a basis problem. This paper shall explore and study the problem.
2. Related work

There are many different definitions of thesaurus, varying from quite modest definitions that focus on the relations between words without stating which kinds of relations that are meant, to such definitions that state more exactly which relations that are concerned. A modest definition is presented by Schutze and Pedersen: “We define a thesaurus as simply a mapping from words to other closely related words”. Furthermore, a thesaurus for a specific field usually includes a relatively complete set of terms in that field collected and organized by domain experts and knowledge organization professionals, and these terms can be used as concept candidates of domain ontology. A thesaurus is a controlled vocabulary that shows relations (e.g. semantic) between terms, which can aid searchers in finding related terms to expand queries.

There are two kinds of multilingual thesaurus construction approaches. One is manual multilingual thesaurus construction, the other one is automatic multilingual thesaurus construction. Manual thesaurus construction is a time-consuming and quite expensive process, and its content is more or less subjective. There is a need of automatically construct thesaurus to improve in time and cost aspects, which can result in more objective thesaurus and is easier to update. At present, the various methods of monolingual thesaurus construction have been recently presented. Many methods are also reported on the thesaurus automation construction. Most methods relied on the similarity among terms. It assumed statistical models of the terms to observe and calculate similarity from texts, which could be available in the literature (McGill, 1979; Fox, 1986; Crouch, 1990; Saiton et al, 1975). In fact, the methods of almost all traditional thesaurus construction were based on mechanically obtained. Most of the mechanical methods relied on the statistical evidence (Srinivasan 1992; Park, Han & Choi,1995). Traditional methods such as co-occurrence analysis, n-gram analysis and tf-idf (term frequency-inverse document frequency) weighting could be used for this purpose. The concept space method for automatic thesaurus construction was developed by Chen, Ng, Martinez, and Schatz. A concept space was defined as a network of terms and weighted associations which can represent concepts (terms) and their associations in the underlying information space. In 2002, Yuen-Hsien Tseng proposed automatic thesaurus generation for Chinese document based on Chinese lexicon to extraction keywords algorithm. In fact, they were a sort of efficient methods for user searching, but they had their limitations for library personal organizing technology literatures, especially technology literatures in the new fields. And dependent on Chinese lexicon for segmentation, it is unfit for many terms that they do not appear at the Chinese lexicon. In addition, many study are not facing Chinese proposes. Because Chinese and the Indo-European languages are totally different linguistics families, we need to develop our study for Chinese thesaurus automation construction. The Chinese thesaurus should be not only suitable for library personnel to organize technology literatures, but also suitable for ordinary users to retrieval technology literatures.

Although there are a number of previous studies regarding the thesaurus automation construction, there is less in China. China thesaurus can be divided into area thesaurus and comprehensive thesaurus. Comprehensive thesaurus is such as the Chinese Thesaurus, E-government Thesaurus and so on. These thesauri are manually finished by domain experts and knowledge organization professionals. For the automatic construction technology, it is mainly tracking the abroad technology and research. Many studies are shortage of a complete system research and development work.

3. Thinking of thesaurus construction and comparing of thesaurus standards

3.1. Thinking of thesaurus

Traditional, manual constructing thesaurus confined the field of the library and information science personal for information organization and classification. When it is applied to the network and digital environment, its own professional, limitations and timeliness make it difficult to be used for all types of users. Through the literature investigation, we think that thesaurus as a term control tool can play a tremendous role, it must realize enlargement of application object, universal of using method and diversification of service function in the digital age. So the meaning of thesaurus construction has changed. The technical connotation of
thesaurus automation construction includes two aspects: one is that automatic construction technology has to overcome the shortcoming of time-consuming and arduous faults of traditional manual edit, to solve bottleneck of experts’ knowledge acquisition. The other one is that automatic construction technology needs overcome existing detection of current retrieval method for digital library users such as mistake examining and wrong matching, to improve retrieval efficiency of literature information resources.

3.2 Similarities and differences between Chinese thesaurus standard and International thesaurus standard

Chinese thesaurus standard (GB13190-91) and international thesaurus standard (ISO2788-1986) are the same as their function. And the relationships among the terms are all defined as equivalence, hierarchical and associative relationships. But there exist some differences about the hierarchical and associative relationships. For example, from the point of view of hierarchy relationship to analysis, it can be seen that the international standard is divided into three kinds, generic relationship, the whole-part relationship and instance relationship. Otherwise, Chinese thesaurus standard is chiefly generic relationship. From the point of view of associative relationships to analysis, the international thesaurus standard is more meticulous than Chinese thesaurus standard. The detailed information may refer GB13190-91 and ISO2788-1986. So our study is mainly automatic building generic relationship and associative relationships among terms.

4. Method of thesaurus automation construction

4.1. Automatic extraction of subject terms

The subject terms are normalized words for the expression of theme in literature indexing and retrieval. Building corresponding candidate sets of subject terms is the premise of realizing automatic extraction. Compared with other data sources, frequency of subject terms presenting in science and technology literature is the highest. So our candidate set of subject terms are from science and technology literature. The paper mainly adopts the method based on linguistics and statistical analysis.

4.1.1. Linguistics analysis on subject terms

We have converted EI thesaurus into Chinese. Through analysis to Chinese thesaurus and EI thesaurus, we can draw some conclusion as following:

In the EI thesaurus and Chinese thesaurus, the length of terms is from 1 to 10 basically. Length in 2-6 words is about 80 percent of the total, length in 1 word is about 10 percent of the total. Length in 7-10 words is about 1 percent of the total. Pu ZHANG pointed that length of Chinese term was in 2-6 words, and was about 76.9% percent of the total. Rong ZHANG also pointed that length of Chinese term was majority in 2-6 word and about 71.723 percent of the total, more than 6 words was only about 0.572 percent of the total after analysing term database containing 328150 terms. So, our statistical results are similar with Pu ZHANG and Rong ZHANG’s statistical results. Therefore we think that length in 2-6 words is as a rule of selection and extraction in follow-up study.

In addition, through sampling, participle tagging and syntax analysis to the Chinese thesaurus and EI thesaurus, we draw the grammar structure of subject terms as selection and extraction rules as follows:

At least one verb, noun or noun component in a subject term;

Last one is verb, noun or noun component in a subject term;

First one is not preposition or quantifier in a subject term;

No conjunctions, pronouns and modal particles in a subject term.

To above conclusions, the paper presents experimental data in section 4.4.
4.1.2. Subject term extraction

The basic of subject terms extraction is realizing to term extraction from technology literatures. At present, the domestic research on Chinese term extraction is opposite less. JiangJun WANG conducted a preliminary research on text term extraction of information field. JianZhou LIU used pure statistical methods to automatic term extraction for open corpora. WenLiang Chen obtained field term words based on bootstrapping methods. Through investigation and study to the domestic existing extraction methods, we found that the method based on statistical information is difficult to solve data sparseness problem, which the calculation of statistical model is mainly dependent on the scale of corpus. So, our study is based on linguistics rules and statistical analysis. Linguistics rules are stated as above, statistical method adopts the Mutual Information calculation method in information theory.

Assume term $T_i = \{t_j\}, t_i = t_1 t_2 ... t_{m-r}, t_j = t_{r+1} ... t_n$, $t_i$ is the word or substring in a term. Then,

$$MI(t_i; t_j) = \log_2 \frac{p(t_i, t_j)}{p(t_i) \times p(t_j)}$$  \hspace{1cm} (1)

$p(t_i)$ is the probability of $t_i$ alone in all documents.  
$p(t_j)$ is the probability of $t_j$ alone in all documents.  
$p(t_i, t_j)$ is the probability of $T$ in all documents.

If $t_i$ and $t_j$ are closely integrated, the value of $MI(t_i, t_j)$ is larger. Otherwise, the value is relatively small. Therefore, we use the mutual information value to measure a term that its internal substring bonding strength.

In order to determine the possibility that terms become subject terms. We draw into positional and frequency information of every terms. To technology literature, the terms that appear at the title, the keywords and abstract are more important than other position. So we calculate the tf-idf value of every term combining position factor $\alpha$. The value of $\alpha$ is different at different positions.

To keywords, calculation formula is  

$$weight = a \times TFIDF_i$$ \hspace{1cm} (2)

To non-keywords, calculation formula is  

$$weight = MI \times a \times TFIDF_i$$ \hspace{1cm} (3)

$$TFIDF_i = \frac{tf_i \times \log(N/n_i)}{\sqrt{\sum tf_j \times \log(N/n_j)}}$$ \hspace{1cm} (4)

$tf_i$ is the frequency of $t_i$ in document $d$.  
$N$ is the number of documents. 
$n_i$ is the number of documents that contains $t_i$.

4.2. Automatic building relationships among terms

In fact, automatically building relationships among terms is a complicated problem, especially to Chinese language. In the paper, we mainly discuss generic relationship and associative relationship. We also think that the methods of distinguishing associative relationship are different to different kinds of literature structure background. We have separately carried on the study of building relationship for two different literature structures. One is journal literatures, the other one is international patent classification. To different literature structure, the method of building generic relationship among terms is consistent, the method of building the associative relationship among terms is somewhat distinguish.
4.2.1. Building generic relationship

In Chinese thesaurus, hierarchical relationship mainly refers to the generic relationship. In the Chinese thesaurus standard and international thesaurus standard, the method of judgment is the same. When terms are all follow “whole” and “part” form, their relationships are generic relationships. For example:

The axletree is a part of needlerollerbearing. The needlerollerbearing pertain to axletree. So the two terms are generic relationship between axletree and needlerollerbearing.

So we firstly look for this generic relationship from terms according to their form.

Then calculate their similarity. It is reference to the formula by proposed Yuan WANG and Zhiqiang WU.

\[
similarity = \alpha \times \frac{\text{sim} \_ \text{number} + \text{sim} \_ \text{number}}{\text{sub} \_ \text{number} + \text{number}} + \beta \times \frac{\sum q\text{sim} \_ \text{number}(i)}{\sum \text{sub} \_ \text{number}(i)} + \frac{\sum q\text{sim} \_ \text{number}(i)}{\sum \text{number}(i)} / 2
\]

\[
isim \_ \text{number} \text{is the number of similar word in a term.}

\text{sub} \_ \text{number} \text{is the number of word in a subordinate term.}

\text{number} \text{is the number of words in a term.}

\sum q\text{sim} \_ \text{number}(i) \text{ is the weighing value of the similar word in a subordinate term.}

\sum \text{sub} \_ \text{number}(i) \text{ is the weighing value of the similar word in a term.}

\alpha, \beta \text{ are the adjustable parameters.}

dp \text{ is the length proportion of a subordinate term and a term.}

4.2.2. Building associative relationship

This concept of associative relationship may be fuzzy, in fact, it has no unified standard for associative terms. So the automatic building is difficult. We made a preliminary study on associative relationship aimed at two different science and technology literatures.

4.2.2.1. Building associative relationship among terms based on the co-occurrence statistical analysis

If two terms often occur together in the same document or the same unit window, we think that meaning of two terms may be associative. We calculated and analyzed co-occurrence information of terms from three windows, title, abstract and document. Then, the similarity was got by calculating the joint probability distribution. Computation formula is as follows:

\[
similarity(A,B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{P(A,B)}{P(A)P(B) + P(A,B) + P(A,B)}
\]

\[P(A,B)\text{ is the co-occurrence probability of term A and B appearing in the same window.}

\[P(A,\overline{B})\text{ is the co-occurrence probability of term A appearing in a window, and term B is not appearing in the window.}

\[P(\overline{A},B)\text{ is the co-occurrence probability of term B appearing in a window, and term A is not appearing in the window.}

\]
4.2.2.2. Building associative relationship based on the literature structure

International patent classification is a kind of special structure of scientific and technological literature. As the most innovative literature, patent literature is a kind of important document resources in the technology literatures. The content of international patent classification is divided into eight categories. According to its structure, classified number is divided into department, categories, small kind, large group and group. In language description, it is more close to the natural language statement.

Through analysis figure 1, we can find that the relationship of same department is a kind of top-down layers. In some sense, the relationship of layers has become a kind of "natural" associative relationship. In addition, the terms of international patent classification have serious data sparseness. So the method of statistical analysis is not applicable to it. We think that linguistics rules and content hierarchical relation are more important for obtaining associative relationship among terms. From the associative relationship among layers, we easy to automatic get the associative relationship among terms. The results of extracting are shown in section 4.4.

4.4. Experimental design and analysis

This study is an exploratory work. At present, we have acquired preliminary research achievements. Our experiment results show that the methods described above can get some beneficial information that can be used for thesaurus automation construction.

Experimental data were from Chinese journal literatures about computer sciences in NSTL and Chinese international patent classification (IPC) in the national intellectual property office site. All data were handled by the Chinese Academy of Sciences's POS Tagger tools ICTCLAS and Stanford POS Tagger tools. Then, we were according to methods described above to carry out our experiments. At last, the paper adopted the statistical method to quantitative analysis experimental results.

First, we verify proposed linguistic rules. The experiment extracted terms automatically according to our linguistics analysis result. We compare the differences in precision of extraction terms between two methods, which one adopted linguistics rules, the other one did not adopt. The statistic results shows in Figure 2. We can find that the number of terms is more than 5-10 percent if we adopt linguistics rules. By sampling, most of new terms are correct. For example, we obtained new terms, 统计机器翻译(Statistical machine translation), 跨语言检索(Cross-language retrieval), 句法分析(Syntax analysis) and so on.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of terms without adopting linguistic rules</th>
<th>Number of terms with adopting linguistic rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPC</td>
<td>37689</td>
<td>41877</td>
</tr>
<tr>
<td>Computer journal literatures</td>
<td>38752</td>
<td>64687</td>
</tr>
</tbody>
</table>

**Figure 2.** The statistical data of extracting terms adopting linguistic rules or not
Exploration and Study of Chinese Thesaurus Automation Construction for Digital Libraries
Wen ZENG, Wen ZENG
Journal of Convergence Information Technology, Volume 6, Number 4. April 2011

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of terms with adopting linguistic rules</th>
<th>Number of terms with adopting linguistic rules and Mutual Information calculation</th>
<th>Number of Subject terms with combining position factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer journal literatures</td>
<td>64687</td>
<td>18753</td>
<td>1900</td>
</tr>
</tbody>
</table>

**Figure 3.** The statistical data of extracting terms in different sections

Figure 3 shows the statistical results of adopting linguistics rules, adopting linguistics rules and MI calculation. The extracted subject terms were obtained at last. We did not deal with IPC corpus because of its data being serious sparse. IPC data are inapplicable to statistical analysis. Figure 4 shows the part subject terms extracted from computer journal literatures. In order to evaluate experimental results, we sample randomly 800 from terms and subject terms. By distinguishing artificially, the precision of terms is about 72.5%, the precision of subject terms is about 51.2 %. Computational formula is as follows:

\[
\text{precision}_{\text{terms}} = \frac{\text{Number}_{\text{accuracy}}_{\text{terms}}}{\text{Number}_{\text{terms}}} \times 100\% \approx 72.5\% \tag{9}
\]

\[
\text{precision}_{\text{subject terms}} = \frac{\text{Number}_{\text{accuracy}}_{\text{subject terms}}}{\text{Number}_{\text{subject terms}}} \times 100\% \approx 51.2\% \tag{10}
\]

**Subject term** | **Weighting values** | **Subject term** | **Weighting values** |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>语义</td>
<td>0.003577167</td>
<td>统计机器翻译</td>
<td>0.001999147</td>
</tr>
<tr>
<td>信息</td>
<td>0.002942094</td>
<td>计算机</td>
<td>0.001536703</td>
</tr>
<tr>
<td>实例</td>
<td>0.002374992</td>
<td>机器翻译</td>
<td>0.001440548</td>
</tr>
<tr>
<td>句法分析</td>
<td>0.00176301</td>
<td>词汇消歧</td>
<td>0.001378814</td>
</tr>
<tr>
<td>检索</td>
<td>0.001686567</td>
<td>软件</td>
<td>0.001188847</td>
</tr>
</tbody>
</table>

**Figure 4.** The part subject terms extracted

Figure 5 and Figure 6 shows the generic relationship and associative relationship extracted automatically. From these results, we can find that the relationships are good except for the number of relationships are larger than our desired results. But from another angle, these relationships can help users expand search range. How to refine these relationships is our future work.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Generic relationships</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>机器翻译</td>
<td></td>
<td></td>
</tr>
<tr>
<td>至汉机器翻译</td>
<td>64.26666666666665</td>
<td></td>
</tr>
<tr>
<td>汉至机器翻译</td>
<td>64.26666666666665</td>
<td></td>
</tr>
<tr>
<td>多语机器翻译</td>
<td>64.26666666666665</td>
<td></td>
</tr>
<tr>
<td>汉法机器翻译</td>
<td>64.26666666666665</td>
<td></td>
</tr>
<tr>
<td>汉蒙机器翻译</td>
<td>64.26666666666665</td>
<td></td>
</tr>
<tr>
<td>多语言机器翻译</td>
<td>59.62406015037593</td>
<td></td>
</tr>
<tr>
<td>多策略机器翻译</td>
<td>59.62406015037593</td>
<td></td>
</tr>
<tr>
<td>多引擎机器翻译</td>
<td>59.62406015037593</td>
<td></td>
</tr>
<tr>
<td>软件</td>
<td></td>
<td></td>
</tr>
<tr>
<td>翻译记忆软件</td>
<td>47.39583333333336</td>
<td></td>
</tr>
<tr>
<td>机器翻译软件</td>
<td>47.39583333333336</td>
<td></td>
</tr>
<tr>
<td>本土化软件工具</td>
<td>45.127819548872175</td>
<td></td>
</tr>
<tr>
<td>全文机器翻译软件</td>
<td>43.41836734693877</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.** The part generic relationship extracted
5. Conclusions

The primary purpose of this study is to explore a method to realize Chinese thesaurus automation construction based on the freely available resources in the PRC’s National Science and Technology Digital Library. This paper analyses the length and formation rules of subject terms in the traditional Chinese thesaurus and the EI thesaurus from the linguistics angle. The key methods and results are presented in the paper. The methods about the automatically extracting subject terms and building relationship among terms are proposed. From above experimental results, we can find that our thinking and methods is valid. Using these methods we can extract terms, subject terms and build relationships among them automatically. But we shall be clearly aware that fully realizing the Chinese thesaurus automation construction is unrealistic and difficult. But exploration and study of automatic construction technology is necessary. By our exploration and study, we can provide an auxiliary constructing tool to help library personnel for building thesaurus quickly, especially aim to some terms in the new domains that they have not any thesauri to refer, for example, life sciences, new energy automobiles and so on. Our study will reduce the cost of manual construction the Chinese thesaurus.

In our methods, there are some defects about dealing with term redundancy and precision, these problems need to be improved. We will combine with syntactic analysis and data mining technology to extend the quantities and improve quality of terms from technology literatures in the future.

6. Acknowledgments

The project of this paper is supported by the China Postdoctoral Science Foundation funded project (Program No: 20100470392). It is also supported by the advanced research of Institute of Scientific and Technical Information of China funded project (Program No: YY-2010019).
7. References


