Chinese Patent Mining Based on Sememe Statistics and Key-Phrase Extraction

Bo Jin[1], Hong-Fei Teng[2,*], Yan-Jun Shi[2], and Fu-Zheng Qu[2,3]

[1] Department of Computer Science, Dalian Univ. of Tech., P.R. China
[2] School of Mechanical Engineering, Dalian Univ. of Tech., P.R. China
[3] Key Laboratory for Precision and Non-traditional Machining Technology, Dalian Univ. of Tech., P.R. China
tenghf@dlut.edu.cn

Abstract. Recently, key-phrase extraction from patent document has received considerable attention. However, the current statistical approaches of Chinese key-phrase extraction did not realize the semantic comprehension, thereby resulting in inaccurate and partial extraction. In this study, a Chinese patent mining approach based on sememe statistics and key-phrase extraction has been proposed to extract key-phrases from patent document. The key-phrase extraction algorithm is based on semantic knowledge structure of HowNet, and statistical approach is adopted to calculate the chosen value of the phrase in the patent document. With an experimental data set, the results showed that the proposed algorithm had improvements in recall from 62% to 73% and in precision from 72% to 81% compared with term frequency statistics algorithm.

Keywords: Patent, Data Mining, Key-Phrase, Sememe Statistics.

1 Introduction

Patent information is one of the most crucial sources of key technologies for industrial research and development. The World Intellectual Property Organization (WIPO) has predicted that judicious use of patent information could reasonably lead to prevention of duplication of research, which could save as much as 60% in time and 40% in funding. On the basis of an estimate by WIPO, patent publications cover approximately 90–95% of the results of scientific research worldwide, probably greater than the percentage that all scientific journals can cover. Because patent authorities periodically publish patent data for public inquiry, enormous data have been accumulated for many years. However, the tools and concepts used in patent research are not on par with that in scientific research. One of the reasons is the lack of key-phrases.

Key-phrase extraction is considered one of the major barriers to text retrieval, especially for Asian languages [1]. It is commonly known as indexing and finding the phrases that are representative of a document [2]. The needs to harness the

* Corresponding author.

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 516–523, 2007.

tremendous amount of information available from Internet and to achieve semantic retrieval have progressively prompted more interest and research to advance the state-of-the-art of key-phrase extraction [3]. Besides, key-phrase extraction is also the sticking point for automatic classification and text data mining application.

Traditionally, key-phrases have been extracted using various information-process related-methods, such as term frequency statistics [4], machine learning [5], and natural language processing [6]. However, only few examples of key-phrase extraction, particularly in full context, are available. Furthermore, most of the available examples may result in inaccurate and partial extraction.

In this article, we first introduce the semantic structure used in our key-phrase extraction algorithm and then provide the details of our algorithm. The main finding in this study is that we realized a semantic understanding algorithm using traditional statistics approach. Finally, we provide and discuss the obtained results and propose a conclusion and some perspectiveideas.

2 Semantic Structure

The knowledge of semantic structure in this study is obtained from HowNet. HowNet [7] is an online common-sense knowledge base that unveils inter-conceptual and inter-attribute relations of concepts as connotations in lexicons of the Chinese and their English equivalents. The philosophy behind the design of HowNet is its ontological view that all physical and nonphysical matters undergo a continual process of motion and change in a specific space and time. The motion and change are usually reflected by a change in state that in turn, is manifested by a change in value of some attributes. In HowNet, a set of sememes, the most basic set of semantic units that are non-decomposable, is extracted from approximately 6,000 Chinese characters. A total of more than 1,400 sememes are found and organized hierarchically. This is a closed set from which all concepts are defined. The set of sememes is stable and robust enough to describe all kinds of concepts by deriving the set of sememes in a bottom-up fashion. The fact that HowNet has more than 65,000 concepts provides a good proof for its robustness.

Table 1. Description of knowledge specimen in HowNet

ID	Phrase	Sememe
061554	gentleman|男人	human|人,family|家,male|男
005241	must必须	{modality|语气}
114646	plagiarism|剽窃	steal|偷,*copy|抄写
011940	outflank|抄袭	attack|攻打,military|军

Table 1 shows the description of knowledge specimen in HowNet. The method of description and the related rules are based on the Knowledge Dictionary Mark-up Language (KDML), which comprises the following components: (i) over 1400 features and event roles, (ii) pointers and punctuation, and (iii) word order. We set up knowledge description model for key-phrase extraction based on KDML.

3 Key-Phrase Extraction Algorithm

As the key-phrase extraction approach needs to access all the available patent specifications, a highly efficient extraction algorithm, both in terms of space and time, must be employed, or else it becomes too large or too slow for practical use. There has been a tremendous amount of researches in this area [4-6]. In this study, we propose a sememe statistics algorithm (SSA) for key-phrase extraction. Our algorithm adopted statistics approach to deal with the above-mentioned semantic structure. We consider key-phrase extraction as a classification task: each phrase in a document is either a key-phrase or not, and the problem is to correctly classify a phrase into one of these two categories.

Generally, the more frequently a word occurs, the more important is. Therefore, the first factor influencing the chosen function of word is word frequency. To this extent, we have used the relative frequency ratio (RFR) [8]. For a given word, RFR is the ratio of the frequency of the word in the document to the frequency of the word in a general corpus. The idea is that words that occur relatively frequently in a certain document compared to how frequently they occur in a general document are more likely to be good key-phrases.

The second criterion we have used to find important key-phrases relies on a measure of semantic similarity derived from the available lexical resources. There are some related research work of concept counting using the WordNet lexical database [9]. As mentioned above, HowNet is adopted as our lexical resource. For HowNet, since sememe is attached to each concept associated with a particular word, we can measure the semantic similarity with the collection of sememes.

```
Initialization:
    DocSet←{w1,w2,…,wn}
    ThresholdVal←0.3
    ChosenNum←5
Main:
While (DocSet≠Ø) do
    w←pop(DocSet)
    fw[w]←fw(w)
    SemSet[w]←{ s1,s2,…,sm }
    while (SemSet[w]≠Ø) do
        s←pop(SemSet[w])
        fs[w][s]←fs(s)
        fc[w][s]←fc(s)
    endwhile
    S[w]←fw[w]; fs[w]; fc[w]
    If S[w]>ThresholdVal then
        w→ChosenSet
    endif
endwhile
KeyPhraseSet←ChosenSet(ChosenNum)
```

Fig. 1. SSA: Sememe Statistics Algorithm

Let d be a patent document. We note that document d consists of the set of words w: $d=\{w_1,w_2,\ldots,w_n\}$ and that word w consists of the set of sememes s: $w=\{s_1,s_2,\ldots,s_m\}$. Thus, $d=\{(s_{11},s_{12},\ldots,s_{1m}), (s_{21},s_{22},\ldots,s_{2m}), \ldots, (s_{n1},s_{n2},\ldots,s_{nm})\}=\{s_1,s_2,\ldots,s_k\}$. We provide a chosen function for this kind of situation. The function $S(w)$ is defined as:

$$S(w) = \alpha \log(fw(w)+1) + \beta \sum_{i=1}^{m} c_i \tag{1}$$

where α and β are the weighting factors. $fw(w)$ is the relative frequency ratio (RFR) of word w. c_i is the center value of sememe.

Figure 1 showed the instantiation of SSA for key-phrase extraction. As in the untimed case, the algorithm is based on a document set, DocSet, containing all the words of a patent document, and a sememe set, SemSet, containing all the sememes of a word.

3.1 Word-Sememe Conversion

Basically, we have a static corpus that can be preprocessed to convert word to sememe. The corpus includes a vocabulary (covered over 65,000 concepts) and a sememe structure (covered over 1400 sememes).

Firstly, the word segmentation has to be done and the part-of-speech should be tagged because there are no word boundaries in Chinese text. In HowNet, the concept of word or phrase and its description form one entry. Therefore, the coherent sememe of each word can be found with the knowledge dictionary. The word is represented by several sememes and the sememe has various descriptive ways based on the knowledge structure. A "type" parametric should then be defined to describe the sememe type after transferring word into sememe. The word and the sememe may frequently occur during document processing. The more frequently it occurs, the more important is the word or the sememe.

3.2 Relative Frequency Ratio

RFR of word is based upon the fact that the significant words will appear frequently in specific collection of document (treated as foreground corpus) but rarely or even not in other quite different corpus (treated as background corpus). The higher of RFR values of the words, the more informative of the words will be in foreground corpus than in background one.

However, selection of background corpus is an important problem. Degree of difference between foreground and background corpus is rather difficult to measure and it will affect the values of RFR of terms. Commonly, large and general corpus will be treated as background corpus for comparison. In this paper, for our foreground corpus (patent document), we select the PFR People's Daily corpus as compared background corpus. The corpus contains newspaper text from People's Daily.

The relative frequency ratio (RFR) of word w is defined as:

$$fw(w) = \frac{f(d(w))/t_d}{f(PFR(w))/t_{PFR}} \tag{2}$$

where $f(d(w))$ is the frequency at which word w appears in the document, t_d is the total number of words in the document. $f(PRF(w))$ is the frequency at which word w appears in the PFR People's Daily corpus, t_{PFR} is the total number of word in the PFR People's Daily corpus.

3.3 Central Value

The central value is used to review the importance of sememe of word in a document. There are two parameters that influence the central value from different aspects: (i) sememe frequency, and (ii) conclusion degree. The central value is defined as:

$$c(s) = \left[\gamma_1 \cdot \log(fs(s)+1)\right] \cdot \left[\gamma_2 \cdot fc(s) + \gamma_3\right] \tag{3}$$

where $fs(s)$ is the sememe frequency, and $fc(s)$ is the conclusion degree. γ_1, γ_2, and γ_3 are the weighting factors.

We believe that the more words a sememe covers in a document, the more important that the sememe would be. Therefore, the most important factor influencing central value is the sememe frequency. Suppose the set of words which associated with sememe s is $\{w_1, w_2, \ldots, w_n\}$. The sememe frequency of sememe s is defined as:

$$fs(s) = \sum_{i=1}^{n} f(w_i) \tag{4}$$

where $f(w_i)$ is the frequency at which word w_i appears in the document.

In HowNet, the sememes are hierarchically organized with tree structure. In this tree structure, the upper sememes may cover more concepts, but sometimes too general. Whereas the lower sememes may focus on the detailed concepts, but sometimes too limited. Therefore, we consider that the crucial sememe should have certain ability of conclusion and should locate in the middle of sememe tree structure. The conclusion degree is used to detect conclusion ability of sememe. Let the child nodes set of sememe s be $\{s'_1, s'_2, \ldots, s'_n\}$, the ancestor nodes set of sememe s be $\{s''_1, s''_2, \ldots, s''_m\}$. Then the whole nodes set associated with sememe s is $\{s'_1, s'_2, \ldots, s'_n, s, s''_1, s''_2, \ldots, s''_m\} = \{s_1, s_2, \ldots, s_k\}$ $(k=n+m+1)$. The conclusion degree $fc(s)$ reflects the relative importance of sememe s in the tree structure. The higher $fc(s)$ is, the more disperse and equal the content covered, and the more conclusive ability the sememe would have. The conclusion degree $fc(s)$ is defined as:

$$fc(s) = 1 - \frac{fs(s)}{\left[\sum_{i=1}^{n} fs(s'_i)\right] \cdot \left[\sum_{i=1}^{m} fs(s''_i)\right] + Max_{i=1}^{k}(fs(s_i))} \tag{5}$$

where $fs(s)$ is the sememe frequency of sememe s.

3.4 Threshold Value

Formula (1) to (5) can be used to calculate the chosen value of each word in document. The chosen threshold value and selection number should then be set up. If the chosen value of a word is larger than threshold value, the word can be added into the chosen key-phrases set. After calculating all the words in the document, we will chose the largest n (n is the chosen number) chosen value words as key-phrases, if the size of chosen key-phrases is larger than the chosen number. In this study, the threshold value is *0.3* and the chosen number is *5*.

4 Evaluation

We carried out an empirical evaluation of SSA using patent documents from the State Intellectual Property Office (SIPO) of P.R. China. Our goals were to assess SSA's overall effectiveness. As is known, there are no author-assigned key-phrases in patent specification. So the current evaluation of key-phrase extraction from patent document is mainly dependent on manual evaluation. Experts in the relative domain are invited to assign key-phrases for patent document. A total of six evaluators responded to our invitation. The evaluators belonged to professors, researchers, or graduate students who are familiar with patent.

We measured key-phrase quality by quality metrics such as recall and precision. Recall and precision have long been used to assess the quality of literature searches. In this study, recall and precision could be used as indicators of the quality of key-phrase extraction. Precision is the proportion of the returned key-phrases that are correct answer; recall is the proportion of the correct answers that are returned.

5 Experimental Results

We selected randomly 200 patent documents from State Intellectual Property Office (SIPO) of P.R. China as our test data. We extracted 921 phrases after performing the aforementioned procedure. Table 2 shows a simple document and the result of key-phrase extraction.

We compared two key-phrase extraction techniques for Chinese patent mining, conventional term frequency statistics method and proposed sememe statistics method. Experimental results are shown in Table 3. The measure to evaluate the key-phrase extraction is recall and precision. Our algorithm was able to recover 73% of the key-phrases with 81% precision. In comparison, the term frequency statistics algorithm obtained only 62% precision when recall is 72%. From Table 3, it can be said that the sememe statistics using HowNet knowledge in the test data is effective in key-phrase extraction.

Table 2. Result of the algorithm

Test Document	**Title:** 薄膜执行机构 **Abstract:** 本发明涉及一种控制阀的薄膜执行机构，其中齿条设置成可以使齿条/齿轮系统的接触点位于实现运动的压力空间的可移动壁的中心连接直线上，所以薄膜本身或薄膜基板上都没有扭转应力作用，实现往复移动的部件的运动的力的方向基本上和运动路径平行而且相互对称。另外，是齿条和齿轮之间的间隙可以通过带有轴承的支撑辊子被基本消除。根据本发明，提供一种可以轻快移动的执行机构，间隙或在活塞-缸系统内的壁摩擦力引起的延迟作用效果可以基本上被消除掉，这种执行机构可以是双作用类型，也可以是单作用类型。 **Assigned Key-Phrases:** 控制阀, 薄膜, 执行机构, 齿轮, 齿条			
Result Key-Phrases	运动	0.907	齿轮	0.382
	薄膜	0.654	齿条	0.382
	执行机构	0.641		

Table 3. Experimental result for different algorithms

Methods	Recall	Precision
Term Frequency Statistics Algorithm	62%	72%
Sememe Statistics Algorithm (SSA)	73%	81%

6 Conclusions and Perspectives

We propose a simple algorithm for key-phrase extraction, called sememe statistics algorithm (SSA). Our algorithm adopted the natural language understanding approach and performed better than the term frequency statistics algorithm in our experiments. We also showed how HowNet can be used by similarity calculation for key-phrase extraction. Experiments on a collection of patent documents showed that our algorithm significantly improved the quality of the key-phrase extraction.

Frequently used key-phrases in a particular domain such as patent would have the additional advantage. This property makes it easier to categorize patent documents using the key-phrase extraction and may benefit to patent search and patent analysis. Furthermore, the proposed method of key-phrase extraction in this paper can be used not only for patent specification, but also for other documents.

Acknowledgments. This work is supported by the National High-tech R&D Program (863 Program) of P.R. China (Grant No. 2006AA04Z109) and National Natural Science Foundation of P.R. China (Grant Nos. 60674078, 50575031).

References

1. Chien, L.F., Pu, H.T.: Important Issues on Chinese Information Retrieval. Computational Linguistics and Chinese Language Processing 1, 205–221 (1996)
2. Schatz, B., Chen, H.: Digital Libraries: Technological Advancements and Social Impacts. IEEE Computer 2, 45–50 (1999)
3. Chen, H., Houston, A.L., Sewell, R.R., Schatz, B.R.: Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques. Journal of the American Society for Information Science 7, 582–603 (1998)
4. Wang, H., Li, S., Yu, S.: Automatic Keyphrase Extraction from Chinese News Documents. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 648–657. Springer, Heidelberg (2005)
5. Freitag, D.: Machine Learning for Information Extraction in Informal Domains. Journal Machine Learning 39, 169–202 (2000)
6. Ong, T.H., Chen, H.: Updateable PAT-Tree Approach to Chinese Key Phrase Extraction using Mutual Information: A Linguistic Foundation for Knowledge Management. In: Proceedings of the Second Asian Digital Library Conference, Taiwan, pp. 63–84 (1999)
7. Dong, Z.D.: Bigger Context and Better Understanding: Expectation on Future MT Technology. In: Proceedings of the International Conference on Machine Translation & Computer Language Information, Beijing, pp. 17–25 (1996)
8. Damerau, F.J.: Generating and Evaluating Domain-Oriented Multi-word Terms from Texts. Information Processing & Management 4, 433–447 (1993)
9. Ji, H., Luo, Z., Wan, M., Gao, X.: Research on Automatic Summarization Based on Concept Counting and Semantic Hierarchy Analysis for English Texts. Journal of Chinese Information Processing 2, 14–20 (2003)