# Study on Hot Topic Discovery from Chinese Texts

Guixian Xu[1], Lirong Qiu[1], Hongfang Liu[2]
[1]Minzu University of China
Beijing, China
[2]Mayo Clinic
Minnesota, USA
xuguixian2000@sohu.com, hol7001@gmail.com

Journal of Digital
Information Management

***ABSTRACT:*** *With the development of information technology, there has been an increased popularity in the use of electronic texts. Topic detection and tracking can identify hot information from isolated texts. Obtaining hot topics has become an important issue in recent years. The combination of statistics and natural language processing was utilized in the current study to discover hot topics from texts. First, the statistics technique was adopted to obtain frequent and high weighted words. Then, the linguistic grammar rules were used to generate the candidate phrases. Finally, the hot phrase topics were obtained based on the weight computation method of phrases. Experiment results showed that the proposed approach was effective in that the extracted topics can express more comprehensive information. This study's results are meaningful in the areas of text classification, text clustering, information retrieval, and construction of high-quality Tibetan corpus.*

**Categories and Subject Descriptors**
**I.2.7 [Natural Language Processing]:** Text Analysis; **I.7 [Document and Text Processing]**

**General Terms:** Information Extraction, Knowledge Management

**Keywords:** Hot Topic Discovery, Topic Detection, Feature Selection, Weight Computation, Information Extraction

**Received:** 17 April 2014, Revised 30 June 2014, Accepted 20 July 2014

## 1. Introduction

With the rapid development of the Internet, text information has become widely used in the creation of web pages. Pieces of relevant Internet information on the same topic are separated in different places and at different times. Achieving a reliable and common cognition with isolated information is difficult [1]. Hence, obtaining vast amounts of information and extracting the hot topics have become important problems requiring immediate solutions. Topic detection and tracking (TDT) satisfies the demand of selecting information on the same topic within a reasonable time. The main task of TDT is to facilitate effective information classification.

Hot topic discovery is the premise and foundation of TDT. A topic is usually defined as an event on which people focus publicly. The hot topics are the highly condensed summary of the texts; they appear, develop, and disappear within a certain period and are characterized as having obvious timeliness. The extraction of the hot topic words is helpful in quickly determining the useful information. Currently, the main technique of hot topic discovery research includes hot TDT [2,3], new event detection (NED) [4], hierarchical topic detection (HTD) [5], and so on. The adopted approaches include the clustering algorithm, natural language processing, and lexical semantic chaining (LSC).

This paper discusses hot topic discovery from Chinese text that combines statistics and linguistic technology. The rest of the paper is organized as follows: The related background is introduced in Section 2, the proposed approach is described in Section 3, the experiment is presented and the results are analyzed in Section 4, and the study is concluded in Section 5.

## 2. Background

The study of hot TDT technology originated in the 1970s, and focused on the research of correlation detection, topic detection, topic tracking, cross-language TDT research, and so on.

NED is an important part of topic detection. Allan [4] and Yang [6] established online identification systems to inspect emerging events. Meanwhile, another work [6] defined a new topic detection task, i.e., HTD. HTD adopted the non-circulation of a root node (directed acyclic graph) to describe the hierarchical topics included. Allan [4] was one of the earliest scholars who used natural language processing (NLP) technology to solve the TDT problem. NLP technology uses the vector space model (VSM) to describe the topics and stories; it provides higher weights to the named entity in the model to imply the NED. Hasan [8] used LSC to describe lexical cohesion, while another study [8] proposed that any word in a sentence could be related to multiple other words in that chain. Morris and Hirst [9] designed the algorithm of automatic construction of LSC based on lexical resources. The lexical chains provided a semantic context for interpreting words, concepts, and sentences [8]. The idea of LSC can be used on topic discovery.

Meanwhile, various dimension-reduction methods have been used to detect topics. For example, latent semantic analysis using singular value decomposition based on VSM [10] compresses a highly dimensional vector space to a lower space. Sriurai [11] applied the topic-model approach to cluster the words into a set of topics, in which words assigned into the same topic must be semantically related. Then these hot topics were used in text categorization. The experimental results showed that the classification model based on hot topics yielded the best performance. Grun [12] used the probabilistic model of R package to obtain the topics in documents. Khodra [13] employed the sentential features to extract the topic sentences of a paragraph. The experiment showed that position and meta-discourse features were important in extracting the topic sentence.

## 2. The Proposed Approach

The use of statistical method based on the hot word extraction technology has become common. However, the expression of the hot word is usually unclear with a single word. The proposed idea of hot topic discovery is to utilize the hot single words and explore them forward and backward. The approach then forms the long two- and three-word pairs, from which hot topic words are discovered more efficiently. For example, "*Beijing*" and " *Olympics*" are two hot words, but when the proposed approach is used, the more useful hot topic, "*Beijing Olympics*," can be generated.

### 3.1 Data Preprocessing

Each Chinese text should be cut into words when the statistical method is employed. ICTCLAS [14] is used to conduct the word segmentation, and the space vector model is adopted to express the document and select the single hot topic words. In this process, the stop words are deleted. Further processing is conducted to remove the words that are not suitable for the topic words based on the parts of speech, such as the following characteristics of a word: /*r* (pronoun), /*q* (quantifier), /*p* (preposition), /*c* (conjunction), /*u* (auxiliary word), /*e* (interjection), /y (modal words), /*o* (onomatopoetic word), /*k* (suffix), /*x* (character string), and /w (punctuation). Finally, term frequency (TF) and document frequency (DF) are used to decrease the dimension.

$D = \{d_1, d_2,\ldots, d_n\}$ is the document collection, $F = \{w_1,w_2,\ldots,w_{|F|}\}$ is the feature set of document collection, and |F| represents the total number of the features. Each document ***d*** is represented as a vector $d = (d_{(1)}, d_{(2)},\ldots, d_{(|F|)})$, and $d_{(i)}$ is the weight of feature $w_i$ in document ***d***. Each feature value in the vector is calculated using the formula $TF * IDF$ (inverse document frequency).

### 3.2 Combination Rules of The Hot Topic Words

The combination rules of the hot topic words are introduced based on the linguistic characteristic of the Chinese text. If a word "*W*" is a single hot topic, moving one step forward or backward around it forms the hot topic of the two-word phase. Moving two steps forward and backward around "*W*" forms the hot topic of the three-word phase. Sample formats would be "*Pre* + *W*, "*W* + *Be*" and "*Pre* + *W* + *Be*," where "*W*" is a word, "*Pre*" is the word before it, and "*Be*" is the word after it. If only the abovementioned method is used, some incorrect combinations may be generated because the grammar rules are not considered. For this reason, the phrase-structure analysis method is used to define the combination rules before the first step, and then the correctness of the word combination is further judged. For example, the three-word phrase made up of "noun + noun + noun" is considered correct, whereas the "*noun* + *adjective* + *adjective*" combination is incorrect. The combination rules of the two- and three-word phrases are shown in Table 1.

### 3.3 Detailed Algorithm

The feature selection is conducted based on the statistical idea of obtaining the word and document frequencies. From the obtained primary candidate words, the candidate word table is built. Considering that the information expression of the individual word may be not perfect, the method combines the word pairs to form phrases and obtain more comprehensive subject information. Grammar rules are also utilized to generate the possible topic phrases. The grammar rules are shown in Table 1. One word is selected in the candidate table, and then the phrase is formed because of the adjacent relation of the words.

Finally, the weights of the phrases are computed, and the hot topics are generated. The detailed algorithm is presented below.

```
Input: a //TF threshold of feature selection
       b //DF threshold of feature selection
       c //the number of hot topic words
Begin:
1). For each document $D_i$ in the text set:
2). Obtain the word table $DW_i$ after the word segmentation of $D_i$.
3). For $DW_i$ , remove stop and single-character words, and exclude the words that are not used as the topic words dpending on the part of speech, such as the words with /r/q/p/c/u/e/y/ o/k/x/w. Then the candidate word table $CW_i$ is generated.
4). Set the document frequency of each word $W$ as 1 in $CW_i$.
5). Divide the document $D_i$ into sentences and obtain the sentence table $DS_i$.
6). For each sentence $S$ in $DS_i$:
7). Look up each candidate word $W$ contained in $CW_i$; if the word $W$ is in the sentence $S$, then update its word frequency (TF).
8). Add the three-word phrase "Pre + W + Be" into $CW_i$ depending on the combination rules, such as nnn/nan/
9). If the three-word phrase is not generated, then add the two-word phrase "Pre + W" or "W + Be" into $CW_i$ depending on the combination rules, such as an/nn/vn/nv/na. Update its frequency (TF).
10). End for
11). End for
12). Merge the words and phrases of all $CW_i$ into the DW_ Base table. Obtain the sum of the TF and DF of the words in DW_Base.
13). Calculate the weight of the word using the formula $TF * IDF$.
14). For multi-word phrase, its weight is equals the sum of the weights of its sub-words belonging to the DW_Base table.
15). For DW_Base, filter out the words whose TF values are less than $a$.
16). For DW_Base, filter out the words whose DF values are less than $b$.
17). Sort the three-word phrases with weight in a descending order.
18). Sort the two-word phrases with weight in a descending order.
19). Sort the words with weight in a descending order.
20). Output $c$ three-word phrases as long-length hot topics.
21). Output $c$ two-word phrases as middle-length hot topics.
22). Output $c$ words as short-length hot topics.
End
```

| Three-word phrase combination rules | Two-word phrase combination rules |
|---|---|
| Noun + noun + noun (nnn) | Adjective + noun (an) |
| Noun + adjective + noun (nan) | Verb + adv (va) |
| Noun + prep + noun (npn) | Verb + noun (vn) |
| Verb + noun + noun (vnn) | Noun + Verb (nv) |
| Noun + noun + verb (nnv) | Noun + noun (nn) |
| Noun + Verb + noun (nvn) | -------------------------------- |
| Adjective + noun + verb (anv) | |
| Noun + conj + noun (ncn) | |

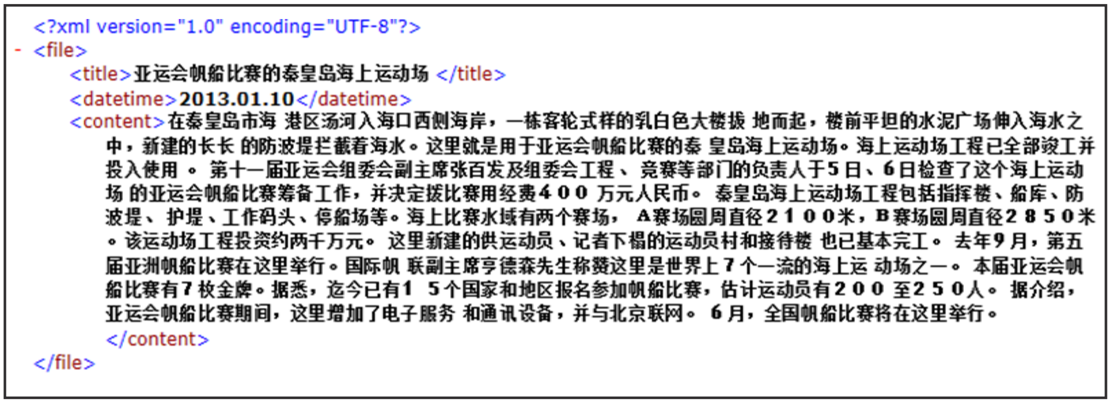
Table 1. Combination rules of the two- and three-word phrases

## 4. Experiment and Results

Fifty XML files about sports are used in the experiment. The format of the XML file includes information, such as title, date, and content. Figure 1 shows an example of the XML file.

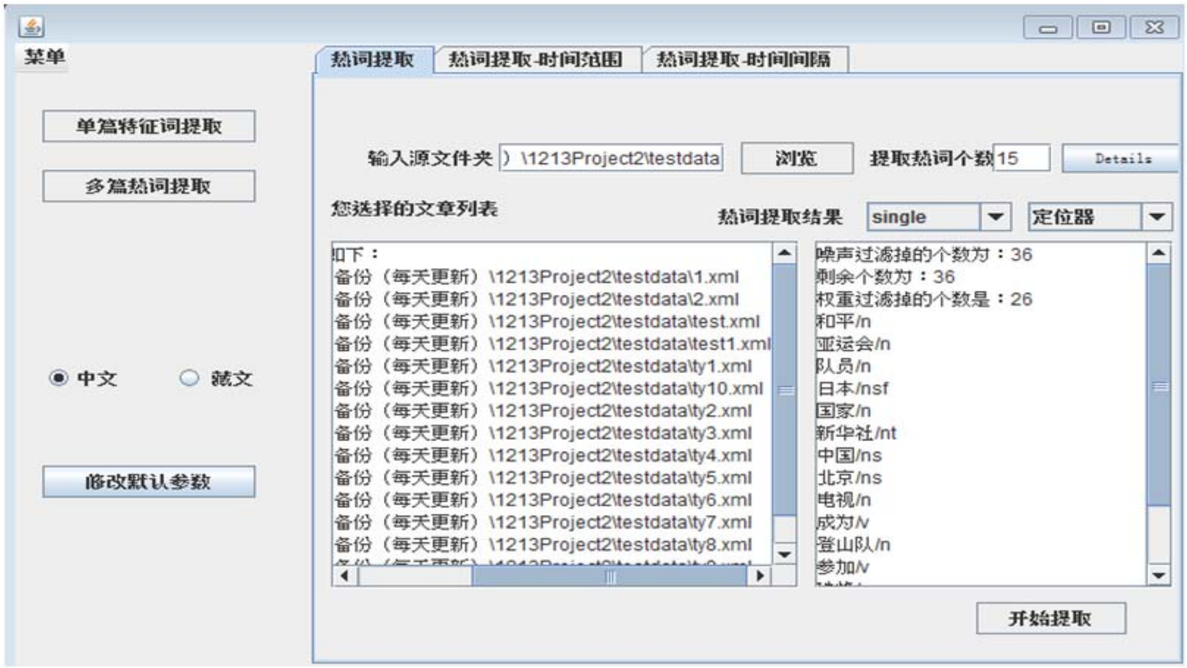
The application system of the Chinese version is developed based on the proposed approach. The following modes are used in discovering the hot topics: (1) extracting hot topic words from the XML file set, (2) extracting the hot topic words from the XML file set based on the date scope, and (3) extracting the hot topic words from the XML file set based on the time interval.

Figure 2 shows the interface of hot topic word extraction



```
  <?xml version="1.0" encoding="UTF-8"?>
- <file>
      <title>亚运会帆船比赛的秦皇岛海上运动场 </title>
      <datetime>2013.01.10</datetime>
      <content>在秦皇岛市海 港区汤河入海口西侧海岸，一栋客轮式样的乳白色大楼拔 地而起，楼前平坦的水泥广场伸入海水之
          中，新建的长长 的防波堤拦截着海水。这里就是用于亚运会帆船比赛的秦 皇岛海上运动场。海上运动场工程已全部竣工并
          投入使用 。 第十一届亚运会组委会副主席张百发及组委会工程、 竞赛等部门的负责人于5日、6日检查了这个海上运动
          场 的亚运会帆船比赛筹备工作，并决定拨比赛用经费４００ 万元人民币。 秦皇岛海上运动场工程包括指挥楼、船库、防
          波堤、 护堤、工作码头、停船场等。海上比赛水域有两个赛场， A赛场圆周直径２１００米，B赛场圆周直径２８５０米
          。该运动场工程投资约两千万元。 这里新建的供运动员、记者下榻的运动员村和接待楼 也已基本完工。 去年9月，第五
          届亚洲帆船比赛在这里举行。国际帆 联副主席亨德森先生称赞这里是世界上７个一流的海上运 动场之一。 本届亚运会帆
          船比赛有７枚金牌。据悉，迄今已有１ ５个国家和地区报名参加帆船比赛，估计运动员有２００ 至２５０人。 据介绍，
          亚运会帆船比赛期间，这里增加了电子服务 和通讯设备，并与北京联网。 ６月，全国帆船比赛将在这里举行。
          </content>
  </file>
```

Figure 1. Example of the XML file



输入源文件夹 ）\1213Project2\testdata

提取热词个数 15

您选择的文章列表

热词提取结果 single 定位器

```
如下：
备份（每天更新）\1213Project2\testdata\1.xml
备份（每天更新）\1213Project2\testdata\2.xml
备份（每天更新）\1213Project2\testdata\test.xml
备份（每天更新）\1213Project2\testdata\test1.xml
备份（每天更新）\1213Project2\testdata\ty1.xml
备份（每天更新）\1213Project2\testdata\ty10.xml
备份（每天更新）\1213Project2\testdata\ty2.xml
备份（每天更新）\1213Project2\testdata\ty3.xml
备份（每天更新）\1213Project2\testdata\ty4.xml
备份（每天更新）\1213Project2\testdata\ty5.xml
备份（每天更新）\1213Project2\testdata\ty6.xml
备份（每天更新）\1213Project2\testdata\ty7.xml
备份（每天更新）\1213Project2\testdata\ty8.xml
```

```
噪声过滤掉的个数为：36
剩余个数为：36
权重过滤掉的个数是：26
和平/n
亚运会/n
队员/n
日本/nsf
国家/n
新华社/nt
中国/ns
北京/ns
电视/n
成为/v
登山队/n
参加/v
```

Figure 2. Hot topic word extraction from the XML file set (Designed in Chinese)

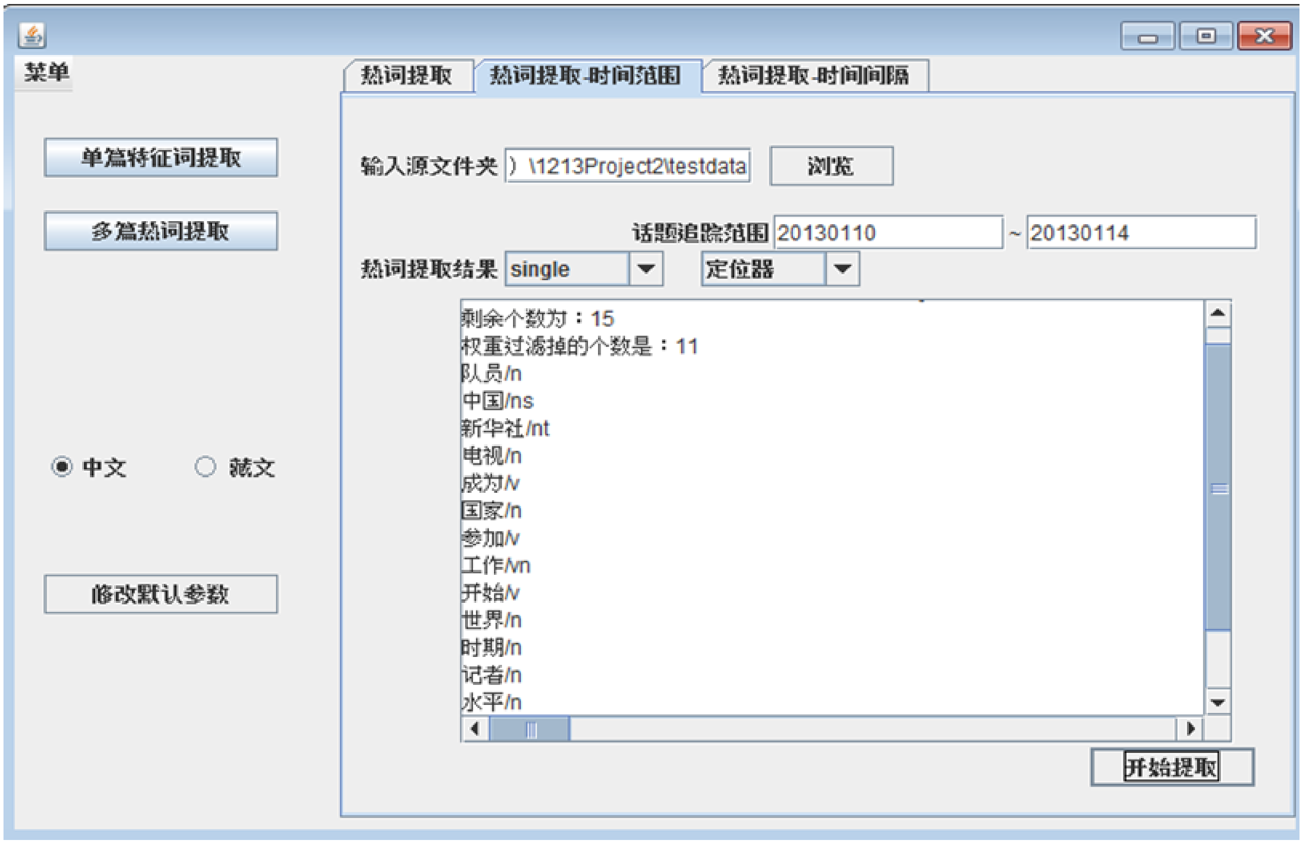
from the XML file set. Figure 3 shows the interface of hot topic word extraction of the XML file set from 2013/01/10 to 2013/01/14 and every day thereafter.

To show the hot topic words, four options are provided for every discovery mode, including "*pre* + *single* + *be*," "*pre* + *single*," "*single* + *be*," and "*single*". The "*pre* + *single* + be" option shows the hot topics of three words, the "*pre* + *single*" and "*single* + *be*" options of two words, and the "*single*" option of one word. The TF threshold is equal to 8 and the DF threshold is equal to 5. The number of hot topics is equal to 15. Figure 5 shows the result interface of the "*pre* + *single* + *be*" option. Figure 3 also shows the result interface of the "*single*" option.

The hot topic extraction is conducted on 50 XML files. Table 2 shows three-word topics and their weights among the top 10, Table 3 shows two-word topics and their weights among the top 10, and Table 4 shows one-word topics and their weights among the top 10.

Tables 2 and 3 prove that two- and three-word topics provide more comprehensive information, respectively, where the topic of each phrase is a combination of the adjacent words. In addition, the co-occurrence of the phrase is higher than the TF threshold, and its DF value is also higher than the DF threshold. The weight of the phrase is higher and ranks among the top 15, indicating that the phrase appears to reach a certain level. The hot topic of the phrase plays an important role in the texts and should frequently appear; this reflects the central idea, which is the advantage of the proposed approach. Table 4 shows that one-word topics are short and incomplete, hence, they cannot fully express the central idea. The abovementioned tables show that if the weight of the one-word topic is higher, the probability of its occurrence in two- and three-word topics would be larger. For example, one-word topics "亚运会" and "登山队" have higher weights, and these words appear in two- and three-word topics because of the cumulative calculation of the long topic weight.

The combination of adjacent words before and after can



输入源文件夹 ）\1213Project2\testdata

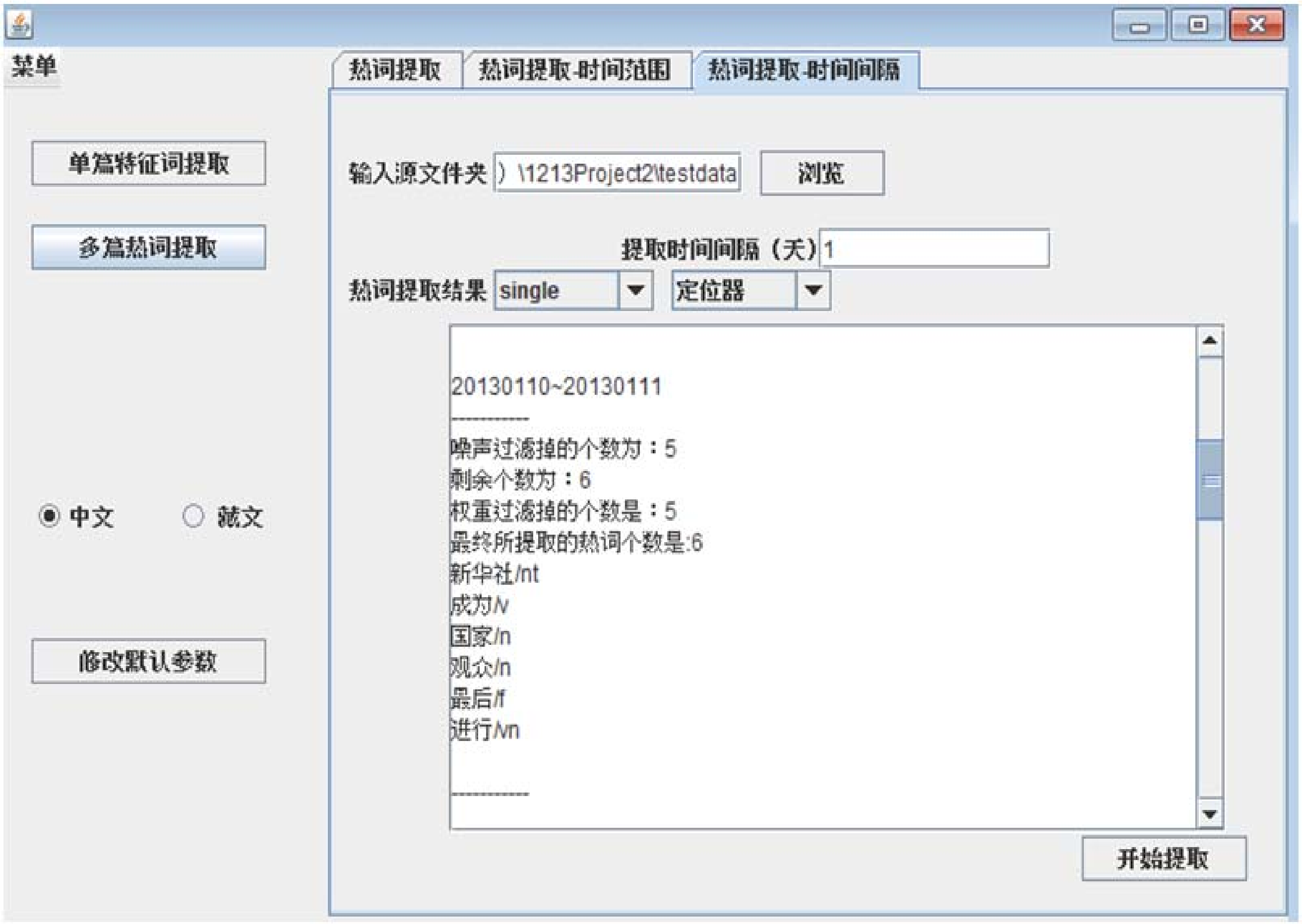
话题追踪范围 20130110 ~ 20130114

热词提取结果 single ▼ 定位器 ▼

● 中文 ○ 藏文

```
剩余个数为：15
权重过滤掉的个数是：11
队员/n
中国/ns
新华社/nt
电视/n
成为/v
国家/n
参加/v
工作/vn
开始/v
世界/n
时期/n
记者/n
水平/n
```

Figure 3. Hot topic word extraction from the XML file set based on the date scope (Designed in Chinese)



输入源文件夹 ）\1213Project2\testdata

提取时间间隔（天）1

热词提取结果 single ▼ 定位器 ▼

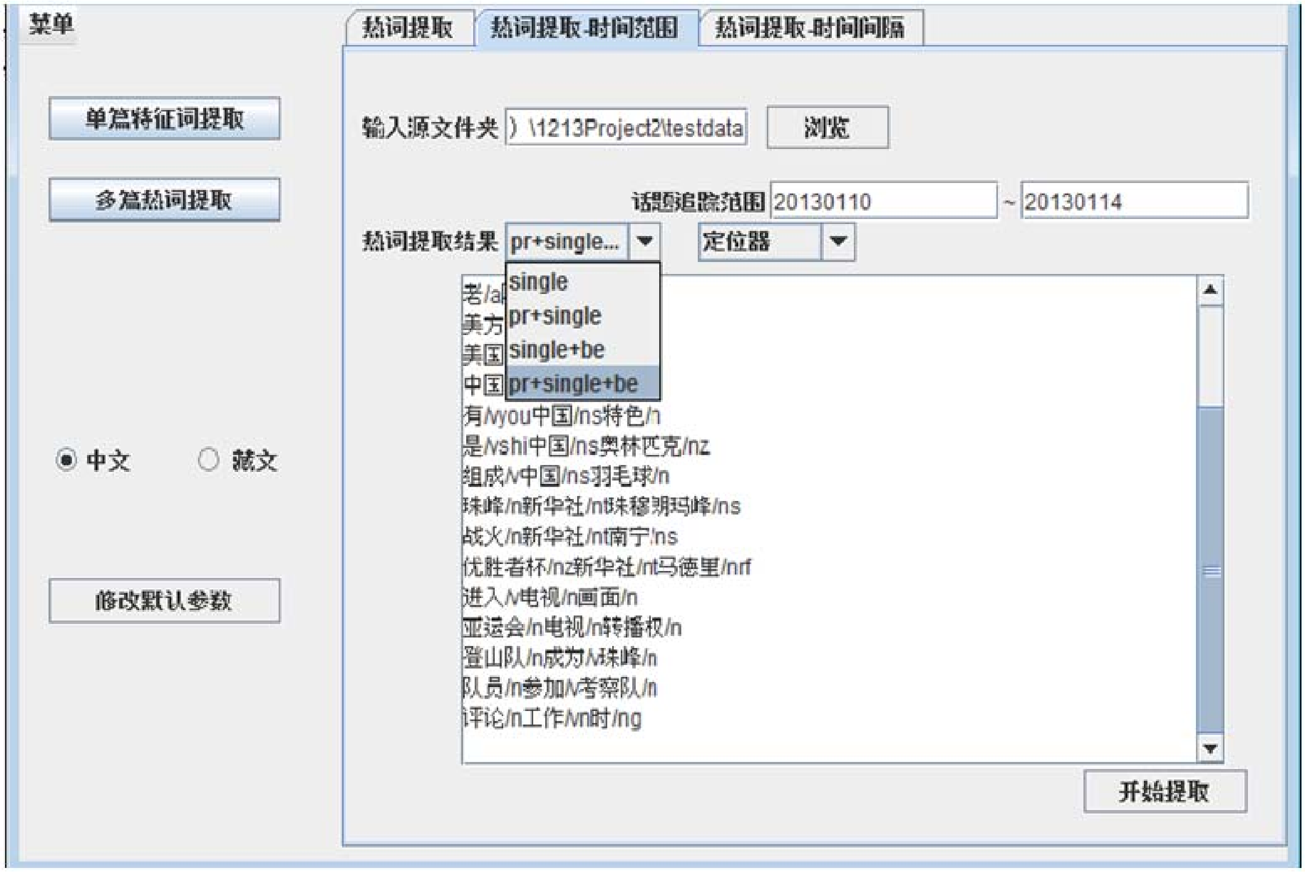
● 中文 ○ 藏文

```

20130110~20130111
------------
噪声过滤掉的个数为：5
剩余个数为：6
权重过滤掉的个数是：5
最终所提取的热词个数是:6
新华社/nt
成为/v
国家/n
观众/n
最后/f
进行/vn


------------
```

Figure 4. Hot topic word extraction from the XML file set based on time interval (Designed in Chinese)



输入源文件夹 ）\1213Project2\testdata

话题追踪范围 20130110 ~ 20130114

热词提取结果 pr+single...

```
老/a
美方
美国
中国
有/vyou中国/ns特色/n
是/vshi中国/ns奥林匹克/nz
组成/v中国/ns羽毛球/n
珠峰/n新华社/nt珠穆朗玛峰/ns
战火/n新华社/nt南宁/ns
优胜者杯/nz新华社/nt马德里/nrf
进入/v电视/n画面/n
亚运会/n电视/n转播权/n
登山队/n成为/v珠峰/n
队员/n参加/v考察队/n
评论/n工作/vn时/ng
```

Figure 5. Hot topics of the “*pre* + *single* + *be*” option (Designed in Chinese)

| Number | Topics | Weight |
|---|---|---|
| 1 | 和平/n 登山队/n 成为/v | 1.6915 |
| 2 | 运动场/n 亚运会/n 帆船/n | 1.6748 |
| 3 | 登山队/n 成为/v 珠峰/n | 1.6505 |
| 4 | 珠穆朗玛/ns 和平/n 登山队/n | 1.5973 |
| 5 | 日本/nsf 队员/n 舟津圭/nr | 1.5149 |
| 6 | 参加/v 北京/ns 亚运会/n | 1.4871 |
| 7 | 队员/n 参加/v 考察队/n | 1.3941 |
| 8 | 用于/v 亚运会/n 帆船/n | 1.3912 |
| 9 | 征服/v 珠峰/n 新华社/nt | 1.3580 |
| 10 | 亚运会/n 电视/n 转播权/n | 1.3120 |

Table 2. Examples of Three-Word topics

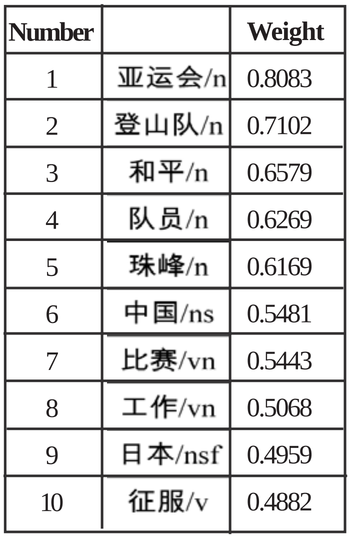
be further improved. Only two- or three-word topics are extracted in the proposed approach. Some longer and meaningful topics can be used in future works on this topic

## 5. Conclusion

Hot topic discovery from Chinese texts is studied in this work to facilitate effective information classification. The proposed approach combines the statistics and linguistic grammar rules to track the hot topics of the texts. The main purpose of the proposed approach is to overcome the shortcomings of the one-word topic, which is unable to clearly express information. The candidate frequent words are obtained based on statistics. Two- and three-word topics are generated by utilizing the grammar rules and the candidate words. The application system of the Chinese version is then developed based on the proposed approach, and more hot topics are achieved by adjusting some relative parameters. The experiment results prove that the proposed method is effective in extracting two- and three-word hot topics that can express stronger and more comprehensive information. In future works, the

| Number | Topics | Weight |
|---|---|---|
| 1 | 和平/n 登山队/n | 1.3682 |
| 2 | 登山队/n 和平/n | 1.3682 |
| 3 | 艺术/n 进行/vn | 1.2861 |
| 4 | 帆船/n 亚运会/n | 1.2605 |
| 5 | 组委会/n 亚运会/n | 1.2318 |
| 6 | 艺术/n 水平/n | 1.2227 |
| 7 | 集资/vi 亚运会/n | 1.2226 |
| 8 | 运动场/n 亚运会/n | 1.2226 |
| 9 | 艺术/n 全国/n | 1.2098 |
| 10 | 亚运会/n 电视/n | 1.1813 |

Table 3. Examples of two-word topics



| Number | | Weight |
|---|---|---|
| 1 | 亚运会/n | 0.8083 |
| 2 | 登山队/n | 0.7102 |
| 3 | 和平/n | 0.6579 |
| 4 | 队员/n | 0.6269 |
| 5 | 珠峰/n | 0.6169 |
| 6 | 中国/ns | 0.5481 |
| 7 | 比赛/vn | 0.5443 |
| 8 | 工作/vn | 0.5068 |
| 9 | 日本/nsf | 0.4959 |
| 10 | 征服/v | 0.4882 |

approach should be expanded and considerably longer and more meaningful topics should be used.

## 6. Acknowledgment

This study was funded by Beijing Social Sciences Funding Program (No. 14WYB040), the National Natural Science Foundation of China (No. 61309012) and the First-Class University and First-Class Discipline Transition Construction Funds of Minzu University of China.

## References

[1] Hu, H. (2014). Research on Ontology Construction and Information Extraction Technology Based on WordNet. *Journal of Digital Information Management*, 12 (2) 114-119.

[2] Holz, F., Teresniak, S. (2010). Towards automatic detection and tracking of topic change. Computational Linguistics and Intelligent Text Processing. Berlin, Germany: Spring-Verlag, p. 327–339.

[3] Song, D., Wang, W., Chen, Y. (2006). Topic Detection and Tracking with a Developed Vector Space Model. *Computer Technology and Development*, 9 (16) 62–67.

[4] Allan, J., Papka, R., Lavrenko, V. (1998). On-Line New Event Detection and Tracking. *In*: Proceedings of SIGIR' 98: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 37–45. NewYork: ACM Press.

[5] Zhao, H., Zhao, T., Zhang, S., Wang, H. (2006). Topic detection research based on content analysis. *Journal of Harbin Institute of Technology*, 10 (38) 1740–1743.

[6] Yang, Y., Pierce, T., Carbonell, J. (1998). A study on Retrospective and On-Line Event detection. *In*: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p. 28–36. CMU, USA: ACM.

[7] Feng, A., Allan, J. (2004). Hierarchical Topic Detection in TDT-2004. *In*: Proceeding of the 7th Conference of Topic Detection and Tracking. http://www.researchgate.net/publication/228682428_Hierarchical _topic_detection_in_TDT-2004.

[8] Hasan, R. (1984). Coherence and cohesive harmony. Flood L, eds. Understanding Reading Comprehension. Newark, Delaware: International Reading Association, p. 181–219.

[9] Morris, J., Hirst, G.. (1991). Lexical Cohesion by The saural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1) 21–48.

[10] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 391–407.

[11] Wongkot Sriurai. (2011) Improving Text Categorization by Using a Topic Model. *Advanced Computing: An International Journal*, 2 (6) 21–27.

[12] Grun, B., Hornik, K. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40 (13) 1–30.

[13] Khodra M. L., Widyantoro D.H., Aziz, E. A., Trilakson, B. R. (2011). Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences. *Journal of Information and Communication Technology*, 5 (1) 17–34.

[14] Huaping, Z., Hongkui, Y., Deyi, X., Qun, L. (2003). HHMM-based Chinese Lexical Analyzer ICTCLAS. In: The Second SIGHAN Workshop on Chinese Language Processing affiliated with 41th Association for Computational Linguistics, p. 184–187.