Research on Tibetan hot words, sensitive words tracking and public opinion classification

Guixian Xu[1] · Changzhi Wang[1] · Haishen Yao[1] · Qi Qi[1]

Received: 5 April 2017 / Revised: 26 June 2017 / Accepted: 28 June 2017 / Published online: 8 July 2017

Abstract The rapid development of Tibetan information technology provides rich resources for Tibetan information processing technology. The construction of Tibetan corpus is the field of Tibetan information processing of basic work. In this paper, we design the system of Tibetan network data collection and web pages preprocessing. It can timely and efficiently access to web resources, and provide a basis for further analysis of Tibetan data. It can establish the Tibetan related corpus, enrich the Tibetan digital resources. It can also alleviate the status of Tibetan corpus data sparse and lack of resources and bring the convenient condition for Tibetan information processing. The hot words reflect the hot spot of Tibetan people's attention in a certain period of time. Firstly, the paper proposes the method for reducing the space dimension of Tibetan news text. It can effectively reduce the complexity of subsequent processing. Secondly, term weighting method is proposed based on improved TFIDF for Tibetan text information extraction. It utilizes the idea that the words of different locations are given different weights to extract the hot words. On sensitive words discovery and classification of public opinion, sensitive thesaurus are collected artificially. Through the sensitive thesaurus comparison, the sensitive words are extracted. Classification of public opinion words is based on the proposed classification formula and the public opinion thesaurus. It will classify one Tibetan text to one public opinion class. In this paper, the software is developed to automatically collect Tibetan web pages from the network, preprocess the web pages, extract the text features and hot words, discover the sensitive words and classify the Tibetan text to one public opinion class. The experiment shows that the Tibetan hot words extraction is effective and Tibetan classification results of public opinion are significant.

Keywords Web crawler · Tibetan hot words · Term weight computing · Sensitive words discovery · Text classification

1 Introduction

China has strongly advocated the construction of information technology so that the Internet propaganda rate in Tibetan areas is increasing year by year. The number of Tibetan Internet users is also growing exponentially. With more and more Tibetan websites appearing, a large number of Tibetan information is generated daily on the Internet. As a vocabulary phenomenon of the Internet age, hot words reflect hot topics and livelihood issues of a country or a region in a period. Hot words have the characteristics of the times and response immediately. Network public opinion is the sum of political beliefs, attitudes, opinions and emotions which Internet users express about phenomenon and problem of government management as well as social life through the Internet. Compared to the reality, the network public opinion spreads faster and wider. How to extract effectively Tibetan information and carry out Tibetan public opinion analysis is very worthy of study.

At present, both Chinese and English information processing techniques have achieved good results, but the researches on Chinese minority languages are in the primary state. In recent years, the increasing of Tibetan and other minority language website provides sufficient information resources for the study of minority languages. Tibetan corpus is important data resource of Tibetan information processing [1], we can summarize, analyze, generalize, and extract relevant knowledge and information from large-scale Tibetan corpus.

✉ Guixian Xu
xuguixian2000@sohu.com

[1] Information Engineering College, Minzu University of China, Beijing 100081, China

Through rapid identification and tracking of hot words [2], we can quickly understand the people feelings, know the social dynamics and development trends. It is helpful to grasp the trend of public opinion quickly and perform the correct guidance of public opinion and propaganda. By analysis of Tibetan public opinion, the time and space distribution of public opinion information can be achieved. Through the deep tracking of hot topics, the source and trend of public opinion can be discovered and it will benefit the government departments to deal with the sudden network of public events.

The paper mainly studies three parts. The first part is to use web crawler to collect web pages from relative Tibetan websites and conduct web pages preprocessing, segmentation as well as remove stopping words. The second part is about the study of hot word extraction. It will calculate feature weight value through the statistical method and finally obtain the rank of hot words for a period of time. The third part is about the study of the sensitive word discovering and public opinion classification based on the sensitive thesaurus.

In the following, we first introduce the related background. The proposed approach is described next. We then present the experimental results and conclude our work.

2 Background

2.1 Corpus collection and preprocessing

In the corpus construction [3] and hot words extraction, the traditional way of corpus construction was through a large number of experts and other human resources to collect, organize and process the data, and finally form the corpus. The original construction method of the corpus is not suitable for large scale data because manual work is too much and the cost is too high. The cycle of the construction is too long so that the corpus cannot be timely updated [4]. As Web 2.0 technology becomes more and more mature, everyone is a content creator. A large scale of language samples on the Internet can be used as the input of the basis corpus. Construction of large-scale corpus based on web can effectively build large-scale raw corpus in the short term and is the foundation of natural language processing research. Liu and others have researched on mining and using the Tibetan web text resources [5]. Their research displayed the distribution and the development of Tibetan web sites and explored the important potential value of Tibetan network corpus.

The web crawler [6] is usually used to crawl web pages from the Internet. The original web pages contain a large amount of information that is not related to the text content, such as the HTML markup language of web pages. This interference information is called noise. Removing web noise is very important for the information preprocessing system. After denoising, the structure's complexity of web page tab can be simplified and the page size can be significantly decreased. Therefore the consumption on the time and space can be reduced in the subsequent processing and the reliability of the information processing results can be improved [7].

In recent years, the technology of web page preprocessing becomes more mature. The web preprocessing technology mainly includes: removing the pages' duplicate and noise.

Effective information can be obtained from the web page through the extracting methods such as visual features [8], DOM tree [9], text features [10] and so on. These methods can remove the web page noise effectively. XML is usually selected to save the useful information extracted from raw corpus. In the specific operation of information extraction of web page, the majority of researchers use DOM4J and JSOUP to preprocess the web page. JSOUP is a Java HTML parser, and it can directly analyze a URL address as well as HTML text content. It provides a labor-saving API to obtain and manipulate data through DOM, CSS, jQuery tools. DOM4J is an open source XML parsing package produced by dom4j.org and is used to read and write XML files. In addition, Dom4J is easy to use and applied to the Java platform. It uses the Java collection framework and fully supports DOM, SAX and JAXP. The performance of DOM4J is excellent. One of the biggest features of DOM4J is the use of a large number of interfaces. It is much more flexible than the previous JDOM in the usage aspect.

About deduplication technology of the web page, Border proposed shingling algorithm and Charikar proposed random mapping method based on the word [11]. These are two current mainstream algorithms: the time complexity of the method shingle is lower, while the accuracy of the algorithm based on random mapping is higher.

2.2 The research status of hot words and public opinions' extraction

The network hot word is a popular vocabulary recently and it reflects the important information of Internet incidents. The related research is still in the initial stage and not enough on the phenomenon of network hot words.

Usually, the extraction of hot words is based on statistical strategy. This strategy is flexible and portable, but it still needs to train a large-scale corpus, and it will generate a lot of useless string affecting accuracy. The whole process needs to split words [12], filter stop words, count frequency of words and do other processing steps. Li Yuqin launched the research of the hot word analysis technology which is based on the data of Internet public opinion [2]. The research is focus on the deep study on hot words' discovery and tracking. The research of the hot words' extraction was mostly based on the frequency of the hot words and their historical frequency fluctuation. Some scholars proposed to set different weight values depending on the words' appearing location as one of the schemes for extracting hot words [13].

Foreign researcher started earlier about the study of the public opinion extraction. Some researches conduct the detection and analysis on the network news, forums, and social media and can grasp the situation of public opinion so as to assist decision-making. The related systems are developed to search and browse the information of public opinion such as SDA (Survey Documentation and Analysis) project of University of California Berkeley and Public Image Monitoring Solution of IBM Company.

During the information processing, the text needs to be represented as a model that the computer can recognize. Vector Space Model (VSM) proposed by G.Salton is widely used in the research of text classification [14,15]. The vector space model procedure can be divided in to three stages. The first stage is to build up the features table of n dimensions after feature selection of the text sets. The second stage is to calculate each weight of the feature appearing in one text based on the features table. The last stage is to use the vector to express every text. For example, the text d is represented as V(d)={w_1, w_2, …, w_n}, where V(d) represents the vector of the text d and w_i represents the weight of the feature $t_i$of the features table in the text d. The weight of the term t_i is usually calculated by term frequency and inverse document frequency (TFIDF).

TFIDF is a numerical statistic that intends to reflect how important a word is in a document [16–18]. It is often used as a weighting factor in information retrieval, text mining, and user modeling. The TFIDF value is proportional to the frequency of the term in the document and inversely proportional to the document frequency of the term. TFIDF algorithm does not reflect the word location information and it is a shortcoming. For the Web document, the calculation method of the weight should reflect the structural characteristics of HTML. The different words in different markers reflect the different importance degree in the article so the weight of the calculation method should also be different. Therefore it is necessary to give different coefficients to different feature words in the web page. It will improve the effect of text representation.

Text classification [19–21] is to build up the classifier according to the predefined topic categories and automatically determines the unknown text to a category. Its study refers to many disciplines such as data mining, computational semantics, artificial intelligence, natural language processing, and information science. At present, statistical classification method, machine learning method, data mining technology and other new technologies are applied to the text automatic classification such as regression model, nearest neighbor classifier, rule learning algorithm, related feedback technology, artificial neural network and so on. These methods learn from the pre-divided classes of text set, acquire the characteristics of each category, automatically generate classification rules, and build a text classifier.

In this paper, the network public opinion [22–24] analysis is divided into two steps. Firstly, conduct the extraction work of sensitive words position by comparing between the text and the corresponding words of the sensitive thesaurus set, and discover the relevant sensitive words. Secondly, get on the public opinion classification on the network text. The text classification is calculated by calculating the frequencies of the public opinion words in the text based on the sensitive thesaurus and computing the similarity between the text and each sensitive class.

3 The proposed method

3.1 Information gathering

Information collection is the first part of the whole project. The extraction of hot words and public opinions needs enough material, while the way of manual acquiring data acquisition cannot meet the needs of research obviously. Therefore, it needs to get a lot of Tibetan corpus by web crawler. Here the open source crawler Crawler4j is selected as the basis to obtain the data. Crawler4j is a lightweight and multi-threaded web crawler. Compared with other open source web crawler, Crawler4j does not need too much complex configuration, and it implementation is simpler. For smaller crawling tasks, using Crawler4j is a better choice.

The designed Crawler is based on Crawler4j and adopts the Breadth-First strategy. The idea is that the web page theme of the initial URL is highly relevant with the theme of web pages in a certain link range and the content is usually latest.

Through the Crawler, web pages are collected from the specific websites and at the same time video, advertising, recruitment information and other irrelevant contents are filtered from the original web pages by the filtering rule. The downloaded pages are stored in HTML document format.

Because the semi-structured HTML documents include much noisy information and cannot be used into the information extraction, HTML web pages need to be preprocessed and stored as structured XML texts.

3.2 Preprocessing of web page

JSOUP is adopted to achieve the useful information in web pages, such as article title, publication time, article content, author and other important information. The noise information on the page is discarded. To improve the preprocessing efficiency, it requires conducting the special analysis for the web page structure of each website and then setting the specific extraction rules' program for different Tibetan website.

3.3 Word segmentation and remove stop words

Word segmentation technology of Chinese text has become mature. The word segmentation technologies include the methods based on dictionary, semantic and statistics. Each method has advantages and disadvantages [25]. With the deeper research of Tibetan information processing, Tibetan text automatic word segmentation technology has make good achievement. Some scholars have realized an automatic Tibetan segmentation scheme based on case-auxiliary words and continuous features [26].

In this paper, we adopt the BCCF algorithm (based on the case auxiliary words and continuous features segmentation algorithm) to deal with Tibetan text segmentation. It is to use the case-auxiliary word and the continuous feature to conduct the word segmentation. It can be divided into the following four steps:

(1) Recognize characters and break sentences: judge whether it's the case-auxiliary word through the length of the character.
(2) Recognize the case-auxiliary word and identify the block (until end of the sentence).
(3) Recognize the words (until end of the sentence): call the dictionary and check continuous feature to match the words.
(4) Conduct the word segmentation (until end of the sentence): conduct the combination which is based on the word's position and the relationship.

The higher the frequency of the word has in the text set, the more likely probability it will be the stop word. So we compute the frequency of each word in the Tibetan text sets and choose some words with the highest frequencies as the candidate stop words [27]. According to the actual situation, The Tibetan experts determine whether each word of the candidate stop words is a stop word. The selected words by the Tibetan experts are stored into stop word list.

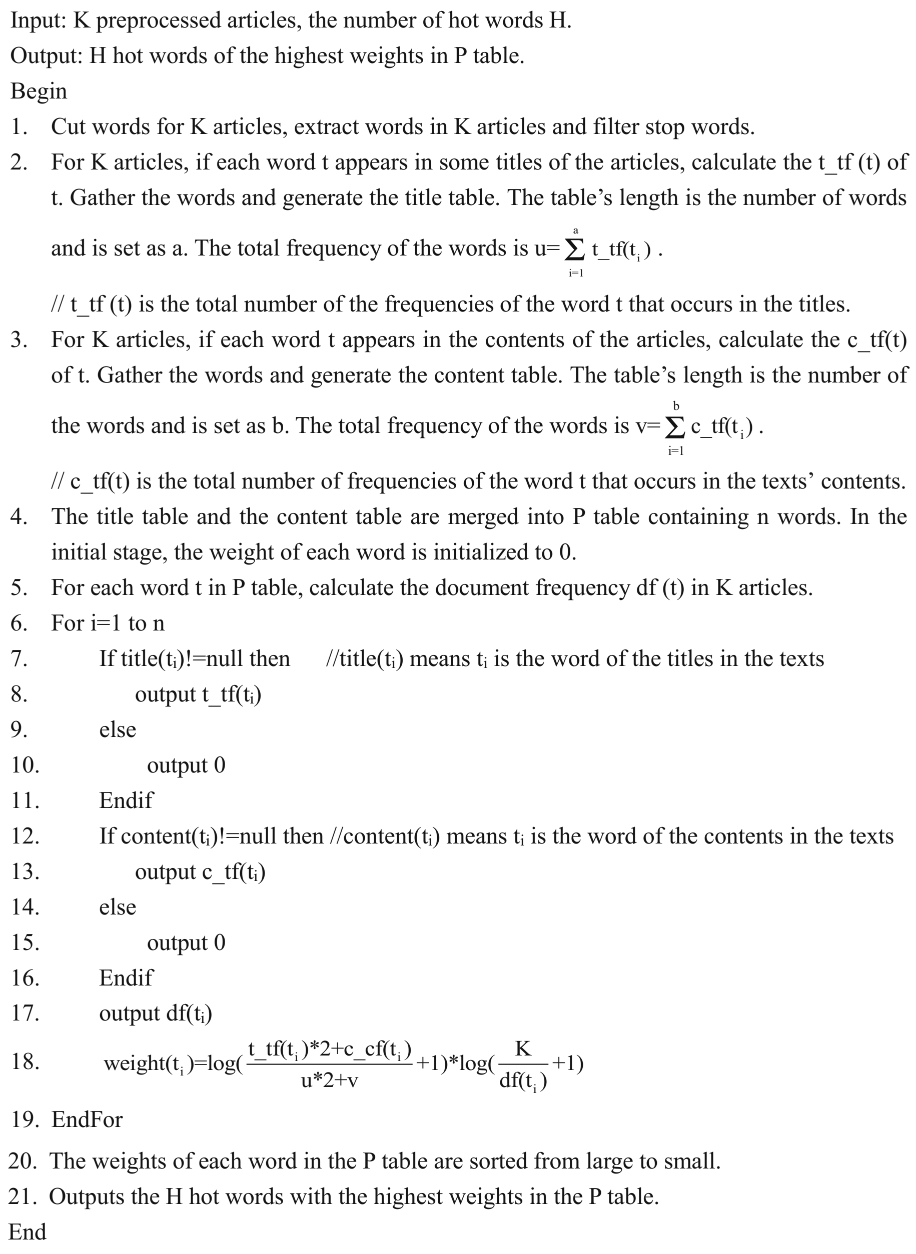
3.4 Hot words extraction

After word segmentation and removing stop words, multi aspect statistics of each word was conducted, which means that it not only needs to count the different frequencies of a word in the different locations in an article, but also needs to count the total appearing frequency of the word in the collected corpus at a certain period of time.

Hot words extraction algorithm is proposed and improved based on the feature weighting method of TFIDF. The algorithm gives the words the different weights according to the different locations in the article, and sets double weight to the word that appears in the title. At the same time, we consider into the number of the total words, the word frequency, the word position and other factors to calculate the weight of the word. Select some words with the highest weight as hot words depending on the weight of each word.

After data collection and preprocessing, data adopt UTF-8 to encode. K text data of some period of time are implemented word segmentation and then stop words are filtered. At last, a large words table P is produced and the number of words in P is n. The P table contains four attributes, the first attribute is the word (word), the second attribute is the term frequency (TF) of the word, the third attribute is document frequency (DF), and the fourth attribute is the weight of the word (weight).

The algorithm for calculating the weights of n words in words table P is as follows:



```
Input: K preprocessed articles, the number of hot words H.
Output: H hot words of the highest weights in P table.
Begin
Cut words for K articles, extract words in K articles and filter stop words.
For K articles, if each word t appears in some titles of the articles, calculate the t_tf (t) of t. Gather the words and generate the title table. The table's length is the number of words and is set as a. The total frequency of the words is $u=\sum_{i=1}^{a} t\_tf(t_i)$ .
    // t_tf (t) is the total number of the frequencies of the word t that occurs in the titles.
For K articles, if each word t appears in the contents of the articles, calculate the c_tf(t) of t. Gather the words and generate the content table. The table's length is the number of the words and is set as b. The total frequency of the words is $v=\sum_{i=1}^{b} c\_tf(t_i)$ .
    // c_tf(t) is the total number of frequencies of the word t that occurs in the texts' contents.
The title table and the content table are merged into P table containing n words. In the initial stage, the weight of each word is initialized to 0.
For each word t in P table, calculate the document frequency df (t) in K articles.
For i=1 to n
    If title($t_i$)!=null then      //title($t_i$) means $t_i$ is the word of the titles in the texts
        output t_tf($t_i$)
    else
        output 0
    Endif
    If content($t_i$)!=null then //content($t_i$) means $t_i$ is the word of the contents in the texts
        output c_tf($t_i$)
    else
        output 0
    Endif
    output df($t_i$)
    $weight(t_i)=log(\frac{t\_tf(t_i)*2+c\_cf(t_i)}{u*2+v}+1)*log(\frac{K}{df(t_i)}+1)$
EndFor
The weights of each word in the P table are sorted from large to small.
Outputs the H hot words with the highest weights in the P table.
End
```

Figure 1 shows the relationship between the tables in the hot words calculation and the attributes of each table.

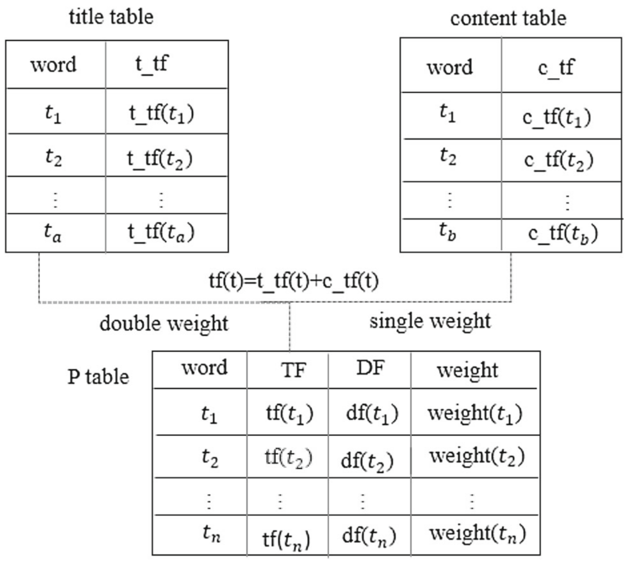
Now simulate one text set and analysis the validity of the hot words' extraction formula through experiments. We let text set have six text's samples (K = 6). The total word frequency of the words in the title is 80 (u = 80) and the total word frequency of the content is 600 (v = 600). Table 1 shows the relevant attributes' values of the seven words and gives the calculated weight value.

Now we analyze the data in the Table 1. The following can be seen:

(1) In the first and second lines of data, the values of title_tf and df are the equal. Because c_tf of t_2 is bigger than t_1 in all contents, the weight (t_2) is significantly higher than weight (t_1).

(2) In the fifth and sixth lines of data, the values of c_tf and df are the equal. Because the t_tf of t_5 is bigger than t_6 in all titles, the weight (t_5) is significantly higher than weight (t_6).

(3) In the sixth and seventh lines of data, the values of t_tf and c_tf are equal. Because the document frequency of $t_5$is bigger than t_6. Weight (t_5) is significantly lower than



title table

word	t_tf
t_1	t_tf(t_1)
t_2	t_tf(t_2)
⋮	⋮
t_a	t_tf(t_a)

content table

word	c_tf
t_1	c_tf(t_1)
t_2	c_tf(t_2)
⋮	⋮
t_b	c_tf(t_b)

word	TF	DF	weight
t_1	tf(t_1)	df(t_1)	weight(t_1)
t_2	tf(t_2)	df(t_2)	weight(t_2)
⋮	⋮	⋮	⋮
t_n	tf(t_n)	df(t_n)	weight(t_n)

Fig. 1 Diagram of each table in the calculation of hot words

Table 1 Test results of weight formula of hot words

No.	Word	t_tf	c_tf	df	Weight
1	t_1	5	20	3	0.042532239
2	t_2	5	25	3	0.04946357
3	t_3	8	15	3	0.043922008
4	t_4	8	15	2	0.055423403
5	t_5	8	25	5	0.041427532
6	t_6	6	25	5	0.037480303
7	t_7	6	25	3	0.052223903

weight (t_6). It meets the rule that the lower the word's document frequency is, the greater the contribution to the information importance.

(4) In the first and third lines of data, df is equal. Both are 3. $t_1$appears 25 times (5 + 20) in 3 articles and t_3 appears 23 times (8 + 15). But t_3 appears 8 times in the titles, the weight of the word in the title is high. So the final weight of t_3 is higher than t_1. It reflects the idea strengthening weight of the words in the title helps to accurately calculate the weight of the word.

(5) Compare the data in the third with the fifth lines. The total term frequencies of t_5 and t_3 are 33 (8 + 25) and 23 (8 + 15). Although the total term frequency of t_5 is 10 more than t_3, due to DF value of t_5 is higher, the t_5's final weight value is smaller than t_3. In the fourth and fifth lines, t_tf is equal, c_tf (t_4) is 15, c_tf (t_5) is 25, the total frequency of t_4 in the contents is 10 less than that of t_5, but the document frequency of t_4 is significantly lower than that of t_5. So weight (t_4) is significantly higher than weight (t_5). It shows that the df attribute is significant to the weight calculation of the word.

1、Yili riots	ཡོང་ནས་ཟིང་འཁྲུག།
2、Urumqi smash	ཨུའུ་ལུའུ་མུའུ་ཆི་བརྒྱབ་ཆོག་དོན་རྐྱེན།
3、East Turkistan riots	ཤར་ཏུ་ཐུ་ཀ་འི་ཟིང་འཁྲུག།
4、Tibetan independence	བོད་རང་བཙན།
5、independence of Taiwan	ཐའི་ཝན་རང་བཙན།
6、terrorist attack	འཇིགས་སྐྲུལ་འཇབ་རྒོལ།
7、camouflage	རྫུ་བ།
8、hostages	མི་གཏའ།
9、rival	འགྲན་ཟླ།
10、antagonist	དགྲ་ཟླ།
11、tactics	འཐབ་རྩལ།
12、death toll	ཤི་བའི་མི་གྲངས།
13、rebellion	དྲག་པོའི་གྱེན་ལངས།
14、rob	འཕྲོག་བཅོམ།
15、barracks	དམག་སྒར།
16、force	དྲག་ཤུགས།
17、invasion	གནོད་འཚེ།
18、weapon verification	མཚོན་ཆ་ཞིབ་བཤེར།
19、expel	མཐར་སྐྲོད།
20、captive	བཟུང་དམག།

Fig. 2 The example of the sensitive word table

3.5 Sensitive information tracking

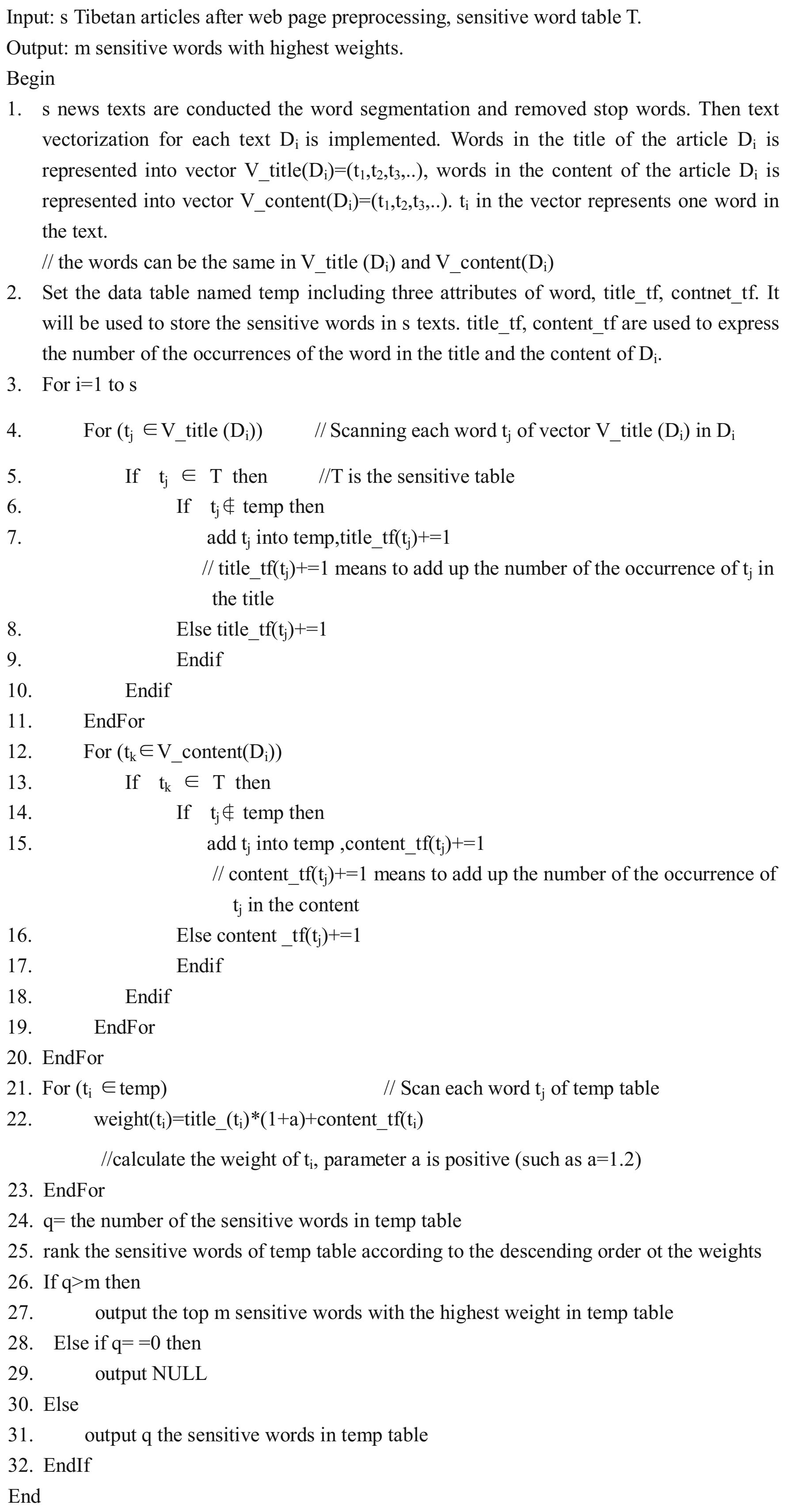
3.5.1 Sensitive word table building

In order to carry out sensitive word tracking, we construct sensitive word table. By means of the artificial selection, 2000 network sensitive words are obtained such as terrorism, violence, religion, social security and so on. The example of the sensitive word table is shown in Fig. 2.

3.5.2 Sensitive word discovering algorithm

After doing word segmentation and removing stop words for the preprocessed Tibetan network data, the following work is implementing to discover the sensitive words. Firstly, conduct data comparison analysis with the sensitive word table and gather statistics frequency of sensitive words appearing in a certain period of time. Secondly, record frequency of sensitive words appearing in each article. Thirdly, calculate the weight of sensitive words and sort the weight of the sensitive words in a certain period of time (such as a day or a week). Finally output the sensitive words with the highest weights, provide a visual monitoring for the public opinion work.

The algorithm for discovering sensitive words is as follows:



```
Input: s Tibetan articles after web page preprocessing, sensitive word table T.
Output: m sensitive words with highest weights.
Begin
1.  s news texts are conducted the word segmentation and removed stop words. Then text
    vectorization for each text $D_i$ is implemented. Words in the title of the article $D_i$ is
    represented into vector V_title($D_i$)=($t_1$,$t_2$,$t_3$,..), words in the content of the article $D_i$ is
    represented into vector V_content($D_i$)=($t_1$,$t_2$,$t_3$,..). $t_i$ in the vector represents one word in
    the text.
    // the words can be the same in V_title ($D_i$) and V_content($D_i$)
2.  Set the data table named temp including three attributes of word, title_tf, contnet_tf. It
    will be used to store the sensitive words in s texts. title_tf, content_tf are used to express
    the number of the occurrences of the word in the title and the content of $D_i$.
3.  For i=1 to s
4.      For ($t_j$ $\in$V_title ($D_i$))      // Scanning each word $t_j$ of vector V_title ($D_i$) in $D_i$
5.          If  $t_j$ $\in$ T then      //T is the sensitive table
6.              If  $t_j$$\notin$ temp then
7.                  add $t_j$ into temp,title_tf($t_j$)+=1
                    // title_tf($t_j$)+=1 means to add up the number of the occurrence of $t_j$ in
                     the title
8.              Else title_tf($t_j$)+=1
9.              Endif
10.         Endif
11.     EndFor
12.     For ($t_k$$\in$V_content($D_i$))
13.         If  $t_k$ $\in$ T then
14.             If  $t_j$$\notin$ temp then
15.                 add $t_j$ into temp ,content_tf($t_j$)+=1
                     // content_tf($t_j$)+=1 means to add up the number of the occurrence of
                       $t_j$ in the content
16.             Else content _tf($t_j$)+=1
17.             Endif
18.         Endif
19.      EndFor
20. EndFor
21. For ($t_i$ $\in$temp)                        // Scan each word $t_j$ of temp table
22.      weight($t_i$)=title_($t_i$)*(1+a)+content_tf($t_i$)
          //calculate the weight of $t_i$, parameter a is positive (such as a=1.2)
23. EndFor
24. q= the number of the sensitive words in temp table
25. rank the sensitive words of temp table according to the descending order ot the weights
26. If q>m then
27.      output the top m sensitive words with the highest weight in temp table
28.  Else if q= =0 then
29.      output NULL
30. Else
31.     output q the sensitive words in temp table
32. EndIf
End
```

In addition to the text and the title, there are some labels stored for each web text for tracking the sensitive words, such as the author, the time of publication, and so on. In the process of tracking the sensitive words, besides the text of the sensitive word information, it can also generate the frequency of sensitive words in the news text, the article title, source site, the total frequency. Through these labels, web sites and other important contents which are related to the sensitive words can be tracked effectively.

3.6 Public opinion classification

The public opinion words are classified into fourteen categories, namely: law, anti-corruption, public health events, education reform, monopoly, department function reform, pornography, social security, social trends, accidents and disasters, the development of network construction, culture, medical and health, the supervision of public opinion. There is a number of public opinion words in each category, and match the network text with the public opinion words of these classes and make a text classification judgment by the similarity algorithm.

The classification algorithm of public opinion for network text is as follows:

```
Input: s preprocessed texts, r public opinion categories vocabularies ($C_1$, $C_2$, $C_3$,…,$C_r$).
Output: the public opinion categories of s texts.
Begin
For i=1 to s
    Conduct word segmentation for the text $D_i$ , remove stop words, form word vector
     V($D_i$).
    For j=1 to r
        Count the matching words’ number m and these words’ total frequency n in
         vector V($D_i$) match with subject thesaurus $C_j$.
        Sim [$D_i$， $C_j$]=(0.7*n+0.3*m)/|$C_j$|.
       // Sim [$D_i$， $C_j$] is similarity of text $D_i$ and $C_j$,
       //|$C_j$| express the words’ number of $C_j$.
    EndFor
    Sort Sim [$D_i$， $C_1$]， Sim [$D_i$， $C_2$]， …， Sim [$D_i$， $C_r$] in a descending order.
    Select the biggest Sim [$D_i$， $C_j$], output $C_j$ the category of $D_i$.
EndFor
End
```

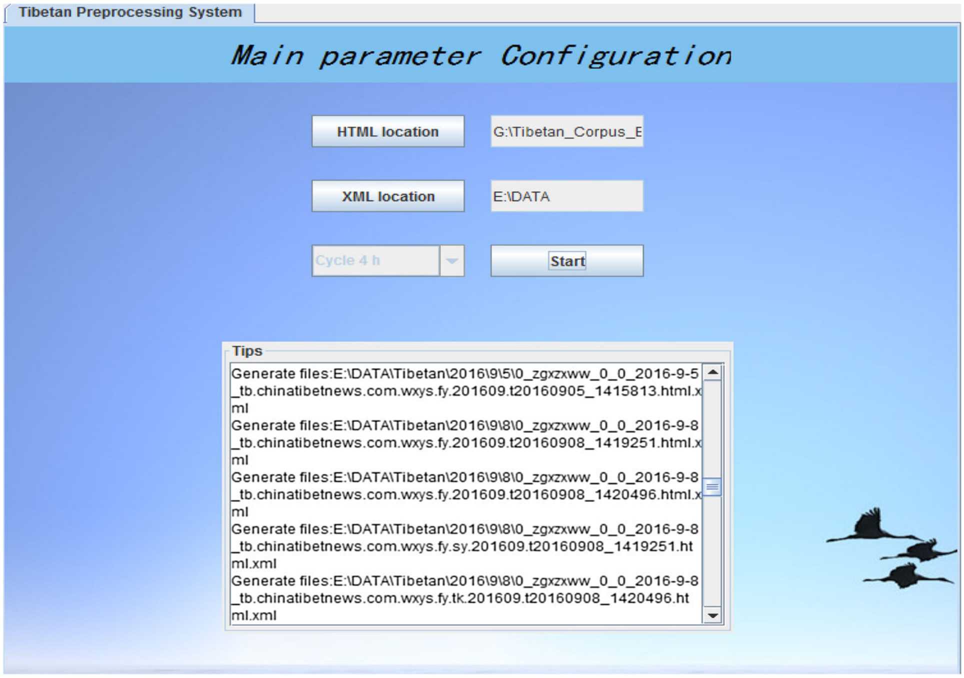
In similarity formula $\mathrm{Sim}[D_i, C_j] = (0.7 \times n + 0.3 \times m)/|C_j|$, similarity computation considers the influence of the matching words’ total frequency n and no repetitive words’ number m between the text D_i and public opinion category C_j. The importance proportion of n and m is 0.7 and 0.3. As for similarity $\mathrm{Sim}[D_i, C_x] = (0.7 \times n_x + 0.3 \times m_x)/|C_x|$ and $\mathrm{Sim}[D_i, C_y] = (0.7 \times n_y + 0.3 \times m_y)/|C_y|$, while $|C_x|=|C_y|$, $m_x= m_y$, if $n_x > n_y$, $\mathrm{Sim}[D_i, C_x]> \mathrm{Sim}[D_i, C_y]$. It indicates that the total frequency number of matching public opinion words between D_i and one category is larger, the probability of being this category is larger. While $|C_x|=|C_y|$, $n_x= n_y$, if $m_x > m_y$, $\mathrm{Sim}[D_i, C_x]> \mathrm{Sim}[D_i, C_y]$. It indicates that the number of no repeat matching words between D_i and one category is larger, the probability of being this category is also larger. While $m_x= m_y$, $n_x= n_y$, if$|C_x| < |C_y|$, $\mathrm{Sim}[D_i, C_x]> \mathrm{Sim}[D_i, C_y]$. It shows that if the number of public opinion words is smaller in one category, the probability of being this category is larger.

4 Experiment and result Analysis

The experiments are developed and implemented on the Java platform. It is an object-oriented visual development language. Java provides a better environment, higher reusability of code, more robust procedures.

4.1 Web Crawler and preprocessing

The designed web crawler is utilized to collect all valid web pages from some web sites. Here crawling range is set to several popular Tibetan web sites such as China Tibet Net (http://www.tibet.cn/), Tibet Xinhua channel (http://tibet.news.cn/), people’s net Tibet channel (http://tibet.people.com.cn/) and Chinese Tibetans Netcom network (http://ti.tibet3.com) and so on. The preprocessing interface is shown in Fig. 3. The crawler got 123636 HTML files. After preprocessing and filtering duplicate files, 102325 XML files are obtained. The saved XML file structure after preprocessing is shown in

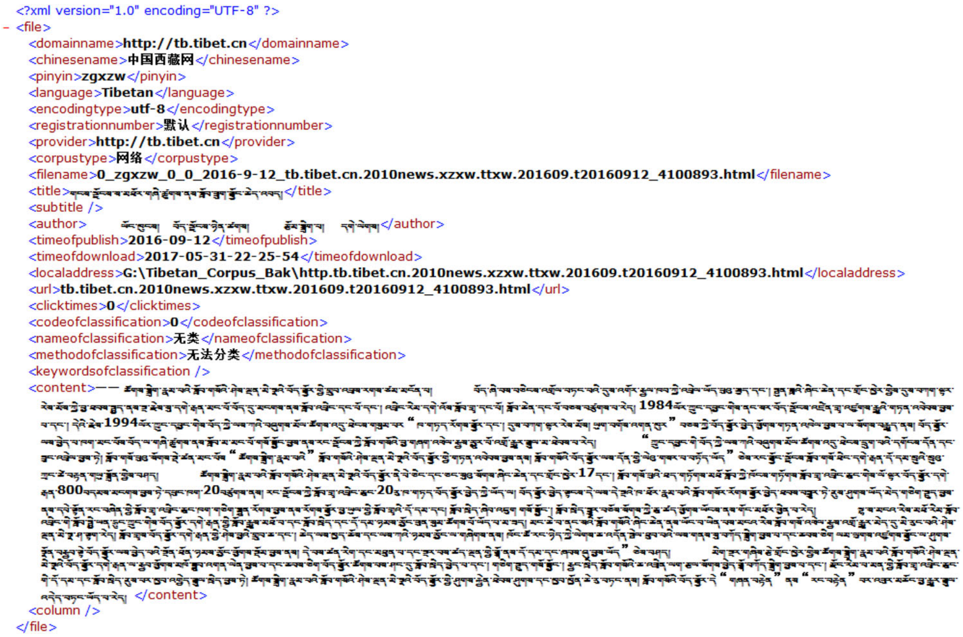


HTML location: G:\Tibetan_Corpus_E

XML location: E:\DATA

```
Generate files:E:\DATA\Tibetan\2016\9\5\0_zgxzxww_0_0_2016-9-5
_tb.chinatibetnews.com.wxys.fy.201609.t20160905_1415813.html.x
ml
Generate files:E:\DATA\Tibetan\2016\9\8\0_zgxzxww_0_0_2016-9-8
_tb.chinatibetnews.com.wxys.fy.201609.t20160908_1419251.html.x
ml
Generate files:E:\DATA\Tibetan\2016\9\8\0_zgxzxww_0_0_2016-9-8
_tb.chinatibetnews.com.wxys.fy.201609.t20160908_1420496.html.x
ml
Generate files:E:\DATA\Tibetan\2016\9\8\0_zgxzxww_0_0_2016-9-8
_tb.chinatibetnews.com.wxys.fy.sy.201609.t20160908_1419251.ht
ml.xml
Generate files:E:\DATA\Tibetan\2016\9\8\0_zgxzxww_0_0_2016-9-8
_tb.chinatibetnews.com.wxys.fy.tk.201609.t20160908_1420496.ht
ml.xml
```

Fig. 3 Main interface of the preprocessing program



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <file>
    <domainname>http://tb.tibet.cn</domainname>
    <chinesename>中国西藏网</chinesename>
    <pinyin>zgxzw</pinyin>
    <language>Tibetan</language>
    <encodingtype>utf-8</encodingtype>
    <registrationnumber>默认</registrationnumber>
    <provider>http://tb.tibet.cn</provider>
    <corpustype>网络</corpustype>
    <filename>0_zgxzw_0_0_2016-9-12_tb.tibet.cn.2010news.xzxw.ttxw.201609.t20160912_4100893.html</filename>
    <title>[illegible]</title>
    <subtitle />
    <author>[illegible]</author>
    <timeofpublish>2016-09-12</timeofpublish>
    <timeofdownload>2017-05-31-22-25-54</timeofdownload>
    <localaddress>G:\Tibetan_Corpus_Bak\http.tb.tibet.cn.2010news.xzxw.ttxw.201609.t20160912_4100893.html</localaddress>
    <url>tb.tibet.cn.2010news.xzxw.ttxw.201609.t20160912_4100893.html</url>
    <clicktimes>0</clicktimes>
    <codeofclassification>0</codeofclassification>
    <nameofclassification>无类</nameofclassification>
    <methodofclassification>无法分类</methodofclassification>
    <keywordsofclassification />
    <content>—— [illegible] 1984 [illegible] 1994 [illegible] 17 [illegible] 800 [illegible] 20 [illegible] 20 [illegible] </content>
    <column />
  </file>
```

Fig. 4 Format of corpus text saving after preprocess

Fig. 4. The annual distribution of the XML files of each Tibetan web from 2014 to 2016 is shown in Table 2. It can be seen that there are 75897 xml files obtained from 2014 to September 12, 2016. Table 2 particularly shows there are 655 XML files from September 5 to September 12, 2016, which will be the basic data of the subsequent experiments. Figure 5 reflects the rapid growth trend of Tibetan web pages year by year. The trend of growth in 2014 and 2015 is particularly obvious.

Experiments show that the use of the web crawler tool can effectively access all the relevant Tibetan web news corpus to achieve real-time pages crawling. Preprocessing function can effectively carry out the structure processing of the web page data, and transform them into the structured XML files for further analysis and processing of data.

4.2 Hot word extraction experiment

The time range of hot words' extraction can be set a day or a week and so on. The data originates from 7 mainstream Tibetan websites. We select news of September 8, 2016 from these Tibetan websites to extract hot words. On this day, total

Table 2 The data distribution of Tibetan web sites after the processing since 2014

Tibetan website	2014	2015	2016 (Due to 9.12)	2016.9.5 ∼ 2016.9.11
tb.chinatibetnews.com	39	1893	5955	375
tb.tibet.cn	4063	2976	1664	40
tibet.cpc.people.com.cn	223	505	379	20
tibet.people.com.cn	2728	7234	5917	69
www.qhtb.cn	11,898	15,740	12,215	136
www.tibetcm.com	381	577	250	5
xizang.news.cn	712	426	122	10
Sum	20,044	29,351	26,502	655
Total	75,897			

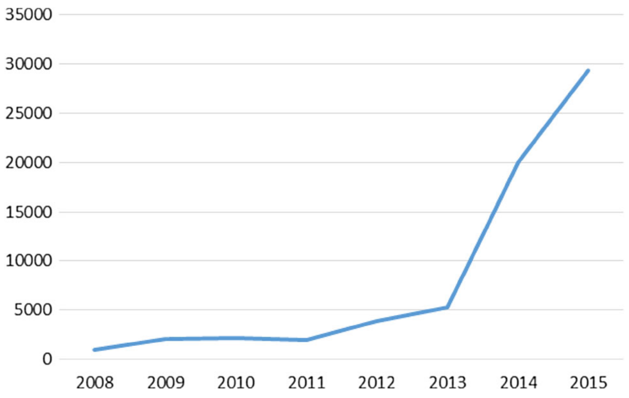


Fig. 5 The annual distribution of XML files after preprocessing

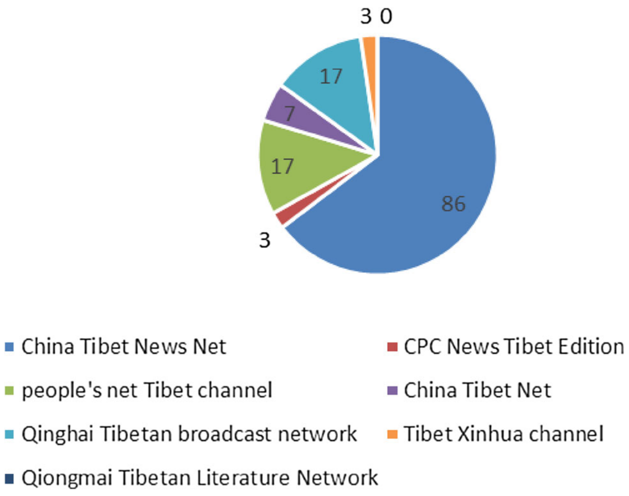


Fig. 6 The distribution of Tibetan news texts on September 8, 2016

133 valid XML texts are used to implement the data analysis after crawling and prepressing. The distribution of 133 texts is shown in Fig. 6.

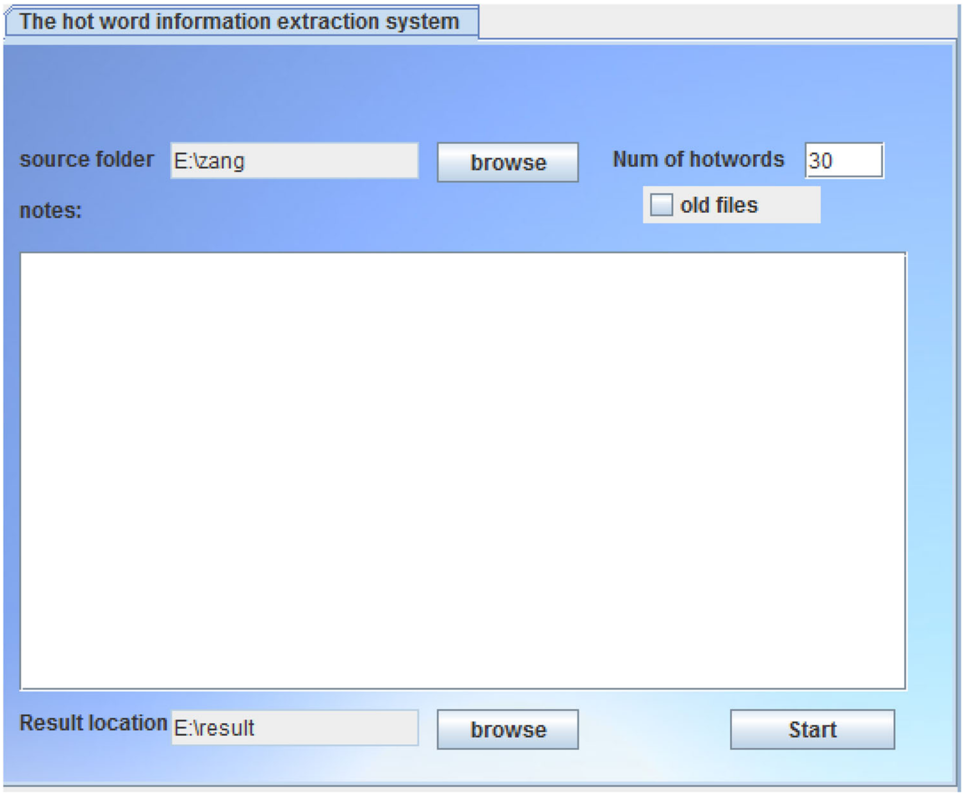
It can be seen from Fig. 6, the number of texts extracted from China's Tibet news website is the largest. It has 86 texts, accounting for 65% of the total news texts. People's net Tibet channel and Qinghai Broadcasting Network both include 17 texts. The proportion is 13%. The number of pages crawled by the Qiong Mai Tibetan literature network is just 0.

The program interface of hot word extraction is shown in Fig. 7. In the hot word extraction tool, "Num of hot words" is used to set the number of the extracted hot words.

The data of September 8, 2016 is used as an example of the hot word extraction of one day. Table 3 shows the top 10 most popular hot words of that day. As can be seen from Table 3, དཔལ་འབྱོར་ is the hottest word in September 8, TF is 241, DF is 33, and the weight is 0.017673684. དཔལ་འབྱོར་, ལམ་ལུགས་, བོད་ལྗོངས་ are top three of the most concerned words, which reflect the focus of people' attention this day. Table 4 shows the top 10 most concerned hot words of one week from September 5 to September 11, 2016. From Table 4, it can be known that རྒྱལ་ཁབ་ is the most popular hot words in this week, TF is 1308, DF is 307, and the weight is 0.012111349. In this week, three hot words with the highest weight are རྒྱལ་ཁབ་， རིག་གནས་, དཔལ་འབྱོར་. It can be seen from Tables 3 and 4 that there are three words in both tables, which indicates that these words may be the subjects of people's constant interest this week.

In Tables 3 and 4, term frequency (TF) shows the times of a word appearing in the text set, Document frequency (DF) value can indicate how many texts include the word. t_tf represents the occurrence number of the word in all news texts' titles, and c_tf represents the frequency of the word that appears in the contents of news texts, and TF value is the sum of t_tf and c_tf. In a text, if term frequency is high and DF value is low, the word will have the strong ability to represent the characteristics of this text. The proposed formula of the weight calculation is more effective in Sect. 3.4. If the stop words are expanded effectively, the extracted hot words will be more representative of the popular events.

The hot word extraction applies the technology of statistics and machine learning to automatically discover and extract the hot spots scattered on the network, and show hot topics to users. The obtained hot information is mainly distributed in the fields of politics, economy, culture and medical care, which reflects the main propaganda direction of the website. It helps people to know the social hot events which happen


The hot word information extraction system

source folder E:\zang

Num of hotwords 30

□ old files

notes:

Result location E:\result

Fig. 7 The program interface of hot word extraction

Table 3 Hot word extraction results on September 8, 2016

No.	Hot words	TF(t_tf,c_tf)	Weight	DF	Chinese
1	དཔལ་འབྱོར་	1241 (6,235)	0.017673684	33	Economics
2	ལམ་ལུགས་	1151 (6,145)	0.013603698	22	System
3	བོད་ལྗོངས་	162 (14,148)	0.011905791	37	Tibet
4	ཁེ་ལས་	1133 (0,133)	0.011530279	22	Enterprise
5	ལག་རྩལ་	1193 (3,90)	0.010318703	13	Technology
6	ཉིན་རེའི་ཚགས་པར་	1184 (0,84)	0.010300626	9	Daily
7	རྒྱལ་ཁབ་	1146 (5,141)	0.010080851	38	Country
8	ལས་ཁུངས་	1182 (1,81)	0.009193922	12	Mechanism
9	བཞོན་མ་	1161 (3,58)	0.008944848	6	Cow
10	ཆོས་ལུགས་	1160 (4,56)	0.008944848	6	Religion

Table 4 September 5–11 hot words extraction results

No.	Hot words	TF (t_tf,c_tf)	Weight	DF	Chinese
1	རྒྱལ་ཁབ་	1308 (46,1262)	0.012111349	288	Country
2	རིག་གནས་	861 (30,831)	0.010387742	178	Culture
3	དཔལ་འབྱོར་	817 (9,808)	0.009687821	176	Economics
4	བོད་ལྗོངས་	695 (63,632)	0.009130896	167	Tibet
5	མི་རིགས་	693 (21,672)	0.008787259	160	Nation
6	དགེ་རྒན་	415 (17,398)	0.008590805	51	Teacher
7	ཀྲུང་གོ་	533 (19,514)	0.00773048	122	China
8	གྲོང་ཁྱེར་	732 (29,703)	0.007713799	232	City
9	ཡུལ་སྐོར་	461 (20,441)	0.007656711	91	Tourism
10	མཚོ་སྔོན་	354 (33,321)	0.006809936	71	Qinghai

in a certain period of time on the Tibetan news web, and provides a theoretical basis for some departments to make relevant policies.

4.3 Discovering experiment of sensitive words

Tracking sensitive word is conducted on the data of a day. According to the sensitive vocabulary table shown in Sect. 3.5.1 and the method of sensitive word extraction shown in Sect. 3.5.2, comparing the news of one day published on the websites (shows in Fig. 6) with the sensitive words' table, the sensitive words of the day can be extracted. Table 5 and 6 show the extracted sensitive word information of September 9 and September 10, 2016.

In Tables 5 and 6, TF attribute includes three numbers. They respectively indicate the term frequency, title's frequency, content's frequency of the word. The term frequency is equal the sum of title's frequency and content's frequency. In the proposed weight formula, parameter a is set 1.2. As can be seen from Table 5, in terms of the data of September 9, two of the most concerned words are ལམ་ལུགས་, སློབ་ཕྲུག. In Table 6,

Table 5 Extracted sensitive words of September 9, 2016

Opinion words	TF	Weight	English
ལམ་ལུགས་	59 (1,58)	60.2	System
སློབ་ཕྲུག་	25 (0,25)	25	Student
ནད་པ་	16 (3,13)	19.6	Patient
སྤྱི་ཚན་	18 (1,17)	19.2	Company
རླངས་འཁོར་	15 (0,15)	15	Automobile
མཉམ་སྦྲེལ་	12 (0,12)	12	Networking
བསྡོམས་རྩིས་	11 (0,11)	11	Statistics
སྲིད་སྐྱོང་	14 (0,4)	4	Ruling
མཚོ་ཐོག་	11 (1,0)	2.2	At sea
ཁེར་རྐྱང་	12 (0,2)	2	Individual

Table 6 Extracted sensitive words of September 10, 2016

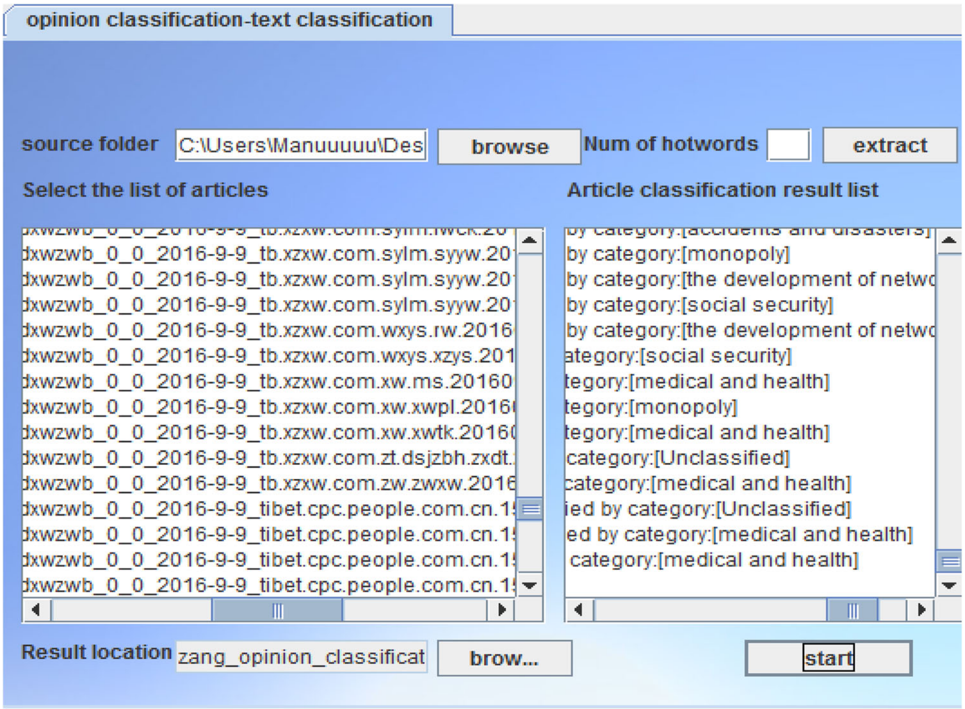
Opinion words	TF	Weight	English
སློབ་ཕྲུག་	37 (0,37)	37	Student
ལམ་ལུགས་	19 (0,19)	19	System
སྤྱི་ཚན་	15 (0,15)	15	Company
སྨན་པ་	12 (0,12)	12	Doctor
ནད་པ་	10 (0,10)	10	Patient
རླངས་འཁོར་	16 (0,6)	6	Automobile
ཁེར་རྐྱང་	12 (0,2)	2	Individual
མ་གནས་	12 (0,2)	2	Cost
མཚོ་ཐོག་	12 (0,2)	2	At sea

three of the most concerned sensitive words of September 10 are སློབ་ཕྲུག་, ལམ་ལུགས་. It can be known that these three words are the focus of people's common concern of these two days. The extracted sensitive words of Tables 5 and 6 can be used to show what people care about during this period of time.

Sensitive word extraction technology can effectively find a variety of sensitive event information and form public opinion of the Tibetan areas. It reflects quickly and directly events' development trends of the Tibetan areas.

4.4 Public opinion classification

The software tool of public opinion classification is based on algorithm of Sect. 3.6 and can automatically classify the preprocessed XML files every day. Public opinion classification software is shown in Fig. 8. The classification software saves the categories of all files to the specified file every day. Influenced by the size of the public opinion vocabulary table, some texts cannot match the public opinion vocabularies, they are marked as the unclassifiable files. For the classified texts, Tibetan experts get on the manual evaluation on the classification accuracy. The result of the evaluation is that the classification accuracy is about 80%. The accuracy of public opinion classification depends on public opinion vocabularies. In terms of the classification performance, public opinion classification experiment is meaningful because it can greatly reduce the effort of the manual classification of text resources and lay foundations for the research work related to the public opinion.


opinion classification-text classification
browse
Num of hotwords
extract
Select the list of articles
Article classification result list
brow...
start

source folder: C:\Users\Manuuuuu\Des

Result location: zang_opinion_classificat

```
dxwzwb_0_0_2016-9-9_tb.xzxw.com.sylm.syyw.20
dxwzwb_0_0_2016-9-9_tb.xzxw.com.sylm.syyw.20
dxwzwb_0_0_2016-9-9_tb.xzxw.com.sylm.syyw.20
dxwzwb_0_0_2016-9-9_tb.xzxw.com.wxys.rw.2016
dxwzwb_0_0_2016-9-9_tb.xzxw.com.wxys.xzys.201
dxwzwb_0_0_2016-9-9_tb.xzxw.com.xw.ms.20160
dxwzwb_0_0_2016-9-9_tb.xzxw.com.xw.xwpl.2016
dxwzwb_0_0_2016-9-9_tb.xzxw.com.xw.xwtk.2016
dxwzwb_0_0_2016-9-9_tb.xzxw.com.zt.dsjzbh.zxdt.
dxwzwb_0_0_2016-9-9_tb.xzxw.com.zw.zwxw.2016
dxwzwb_0_0_2016-9-9_tibet.cpc.people.com.cn.1
dxwzwb_0_0_2016-9-9_tibet.cpc.people.com.cn.1
dxwzwb_0_0_2016-9-9_tibet.cpc.people.com.cn.1
dxwzwb_0_0_2016-9-9_tibet.cpc.people.com.cn.1
```

```
by category:[accidents and disasters]
by category:[monopoly]
by category:[the development of netwo
by category:[social security]
by category:[the development of netwo
ategory:[social security]
tegory:[medical and health]
tegory:[monopoly]
tegory:[medical and health]
category:[Unclassified]
category:[medical and health]
ied by category:[Unclassified]
ed by category:[medical and health]
category:[medical and health]
```

Fig. 8 Main interface of public opinion classification

5 Conclusions

In this paper, the designed web crawler can collect the web pages from the mainstream Tibetan websites automatically and obtain the relevant Tibetan corpus. The collected web pages are converted into the structured files through the relevant web information preprocessing technology. After conducting word segmentation for the texts, removing stop words processing technology, the term frequencies and the positions of the words are counted. DF values of the words are also calculated. The weights of the words are computed based on the proposed algorithm value. Then the hot words having the high weights are selected. The extraction method of the hot words weighs many factors and is effective to find the important information.

For the sensitive words extraction, firstly collect the sensitive thesaurus. Then extract the sensitive words from the text set and track the network texts. For the public opinion processing technology, firstly classify the public opinion thesaurus. Then carry out the classification of the public opinion on the network text by using thesaurus. Experiments show that tracking sensitive Tibetan words and classifying the web texts to public opinion categories are valuable.

This paper provides the study on the extraction of hot words based on the Tibetan web texts. Because of the lack of unified evaluation criteria for extracting network hot words, it causes that the accuracy of hot word recognition cannot be evaluated. The research doesn't take the historical frequency fluctuations of the hot words into consideration. The future research will focus on it. Sensitive words and public opinion words still need to be expanded so that the accuracy of sensitive words' extraction and the classification of public opinion can be improved effectively.

Through the information extraction research of this paper, the focus information can be obtained which people pay attention to in a period of time. Public opinion classification can provide the analysis basis for the decision makers to grasp the public opinion dynamics and offer the service for the relevant departments of the state. This paper will be helpful to the related research work about the Tibetan text information processing.

Acknowledgements This work was supported by the Beijing Social Science Foundation (No. 14WYB040), First class university, First class discipline construction funds of Minzu University of China (No.2017MDYL12), the National Natural Science Foundation of China (No. 61309012), the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No. 2014BAK10B03).

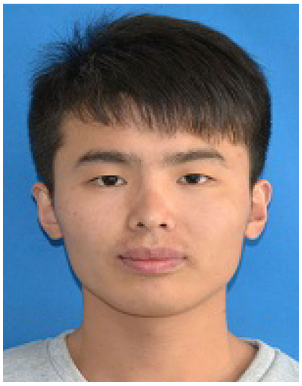
References

1. Gao, D.G., Guan, B.: Retrospect on the development of Tibetan information processing technology. J. Tibet Univ. **24**(3), 18–27 (2009)
2. Li, Y.Q., Sun, L.H.: Hot-word detection for internet public sentiment. J. Chin. Inf. Process. **25**(1), 49–53 (2011)
3. Gao, D.G., Tashigyal, Zhao, D.C.: Data analyses of large basic Tibetan corpus. J. Northwest Univ. Natl. **34**(92), 46–51 (2013)
4. Li, P.F., Zhu, Q.M., Qian, P.D.: Construction approach of large-scale corpus based on web. Comp. Eng. **34**(7), 41–46 (2008)
5. Liu, H.D., Nuo, M.H., Ma, L.L.: Mining Tibetan web text resources and its application. J. Chin. Inf. Process. **29**(1), 170–177 (2015)
6. Yang, D.Z., Zhao, G., Wang, T.: Application of WebCrawler in information search and data mining. Comput. Eng. Des. **30**(24), 5658–5662 (2009)
7. Yang, L., Geng, X., Liao, H.: A web sentiment analysis method on fuzzy clustering for mobile social media users. Eurasip J. Wirel. Commun. Netw. **2016**(1), 1–13 (2016)
8. Wu, Q., Yang, X., Zhao, Z.X.: Web information extraction based on visual characteristics. In: Symposium of the Sixth China Conference on Information Retrieval (2010)
9. Zhang, R.X., Song, M.Q., Gong, Y.L.: Parsing DOM tree reversely and extracting web main page information. Comput. Sci. **38**(4), 213–215 (2011)
10. Hu, J.D.: Research on Web News Extraction and Duplicates Elimination. Zhejiang University, Hangzhou (2011)
11. Ma, C.Q., Mao, X.G.: Research on near-duplicate detection algorithm shingling and simhash. Comput. Digit. Eng. **39**(1), 15–17 (2009)
12. Kang, C., Jiang, D., Long, C.: Tibetan word segmentation based on word-position tagging. In: 2013 International Conference on Asian Language Processing (IALP), pp. 239–242. Urumqi (2013)
13. Jin, Z.: A method of intelligence key words extraction based on improved TF-IDF. J. Intell. **4**, 028 (2014)
14. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
15. Becker, J., Kuropka, D.: Topic-based vector space model. In: Proceedings of the 6th international conference on business information systems, pp. 7–12 (2003)
16. Aizawa, A.: An information-theoretic perspective of tf-idf measures. Inf. Process. Manage. **39**(1), 45–65 (2003)
17. Wu, H.C., Luk, R.W.P., Wong, K.F., et al.: Interpreting tf-idf term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**(3), 13 (2008)
18. Shi, C.Y., Xu, C.J., Yang, X.J.: Study of TFIDF algorithm. J. Comput. Appl. **26**, 167–170 (2009)
19. Cao, H., Jia, H.: Tibetan text classification based on the feature of position weight. In: 2013 International Conference on Asian Language Processing (IALP), pp. 220–223. Urumqi (2013)
20. Jiang, T., Yu, H.Z., Zhang, B.: Tibetan text classification using distributed representations of words. In: 2015 International Conference on Asian Language Processing (IALP), pp. 123–126. Suzhou (2015)
21. Kim, S.B., Han, K.S., Rim, H.C., HyonMyaeng, S.: Some effective techniques for Naive Bayes text classification. IEEE Trans. Knowl. Data Eng. **18**(11), 1457–1466 (2006)
22. Liu, W., Song, Z.: Design and implementation of an internet public opinion monitoring system. In: 2014 International Conference on security, pattern analysis, and cybernetics (SPAC), pp. 114–118. Wuhan (2014)
23. Guo, K., Shi, L., Ye, W., Li, X.: A survey of internet public opinion mining. In: 2014 International Conference on progress in informatics and computing (PIC), pp. 173–179 Shanghai (2014)

24. Li, X., Gao, L.: The design and implementation of an internet public opinion monitoring and analyzing system. In: 2013 International Conference on Service Sciences (ICSS), pp. 176–180. Shenzhen (2013)
25. Mo, J.W., Zheng, Y., Shou, Z.Y., Zhang, S.L.: Improved Chinese word segmentation method based on dictionary. Comput. Eng. Des. **34**(5), 1802–1807 (2013)
26. Chen, Y.Z., Li, B.L., Yu, S.W., Lan, C.J.: An automatic Tibetan segmentation scheme based on case-auxiliary words and continuous features. Appl. Linguist. **1**, 75–82 (2003)
27. Zhu, J., Li, T.R.: Research on Tibetan stop words selection and automatic processing method. J. Chin. Inf. Process. **29**(2), 125–132 (2015)

Guixian Xu is a Vice Professor, College of Information Engineering, Minzu University of China, Beijing, China. Scientific activities: Artificial Intelligence, Data Mining.

Changzhi Wang is a Master student in Software Engineering, College of Information Engineering, Minzu University of China, Beijing, China. Scientific activities: Text Mining.



Haishen Yao is a Master student in Software Engineering, College of Information Engineering, Minzu University of China, Beijing, China. Scientific activities: Data Mining, Natural Language Processing.

Qi Qi is a Master student in Software Engineering, College of Information Engineering, Minzu University of China, Beijing, China. Scientific activities: Data Mining, Machine Learning.