http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5505-5511, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5505.5511

Hot Word Extraction of Tibetan Internet Public Opinion

[1]Sun Yuan and [1]Guo Wenbin
[1]School of Information Engineering, Minzu University of China,
[1]Minority Languages Branch, National Language Resource and Monitoring Research Center,
Beijing, 100081, China

Abstract: Hot word often reflects the hot topics at a particular time. In this study, we research on the key techniques of Tibetan hot word extraction. Through analysis the dynamic Tibetan web data in 2012, we mainly discuss the extraction method of the new hot words and hot related words. (1) Using information entropy and vector space module similarity calculation to extract/filter Tibetan new hot words, (2) Using a frequency-position weighing algorithm to compute weight of word and combining entropy, 3σ criterion and variance to extract and track the Tibetan hot words and (3) Constructing word co-occurrence model to extract and analysis hot related words. Finally, the experimental results prove the method is effective.

Key words: Hot word, frequency-position weighting algorithm, co-occurrence model

INTRODUCTION

With the rapid development of internet, almost 500 million persons make comments, participate in the topic discussion through the forum, blogs and micro-blog in the internet. Internet is becoming the primary place for generation and dissemination of public opinion gradually which plays an increasingly important role in the social life.

As we know, there are 55 minorities and more than 100 kinds of minority languages in China. So far, around 40 minority languages have their own written letters. Mongolian, Tibetan and Uighur have large number of users. Tibetan belongs to the Sino-Tibetan Language Family. The word is formed by superimposing letters. And there is no specific segmentation symbol between the words or sentences. In recent years, Tibetan language webs are rapidly increasing and the number of Tibetan internet users is also fast growing. More and more users express their emotions, attitudes, opinions and demands by internet. And all of these have formed the Tibetan internet public opinion which reflects the social phenomena.

While internet is providing the convenient for people, it also provides some harmful information which posed a serious threat on the national security, such as "Falun Gong", "Tibet separating". At present, the government is still use the manual mode to monitor public opinion. Therefore, how to find hot topics and make data analysis synchronously is a key problem to be solved.

Nowadays, some universities or research institutes have carried out much research work on public opinion monitoring and achieved many fruitful results. The Defense Advanced Research Projects Agency (DARPA) has released a "Deep Exploration and Filtering of Text (DEFT)" program in 2012. This program is expected to provide the capability to identify and interpret both explicit and implicit information from highly ambiguous and vague narrative text and integrate individual facts into large domain models for assessment, planning and prediction. In China, People's Daily Press, Founder Company, Tianya Public Opinion, Public Opinion Research Center of Beijing Jiaotong University are committed to Chinese public opinion monitoring research (Zheng *et al*., 2007, 2012; Yang *et al*., 2010).

With the development of information technology, Tibetan information processing has achieved fruitful results. Tibet University, Northwest University for Nationalities, Qinghai Normal University (Cai, 2009), Minzu University of China (Sun *et al*., 2009, 2010; Dai, 2010) and other research institutes have had good study on the Tibetan information processing (Jiang, 2006). It has provided the basis of monitoring public opinion. Currently, it mainly focuses on monitoring language forms, such as character and word statistics, little analysis for text content.

As the previous study, some hot word extraction methods have been proposed which are mainly based on machine learning and statistics information methods, such as support vector machine or maximum entropy

Corresponding Author: Sun Yuan, School of Information Engineering, Minzu University of China, 100081, Beijing, China

(Tomokiyo and Hurst, 2003; Li, 2004; El-Beltagy and Rafea, 2009; Wang *et al*., 2010). Yang (2008) proposed a joint weight extraction method which considers the frequency, part of speech and semantic relationships of words (2008). Reference (Zheng and Lu, 2005) proposed a method which adopts a non-linear function and comparisons.

Xue *et al*. (2011) proposed a dynamic text model dynamic burst vector space model which can describe the dynamic attributes of text efficiently. Meanwhile, a public opinion analysis system, with a kind of high statistics efficiency improved LC frequent pattern mining algorithm data flow analysis text clustering method is designed to analyze the hot spots (Chen *et al*., 2012). Li Yuqin *et al*. processed deep research for hot word discovering and associating technique. In the phase of word discovering, they utilize named entity recognition techniques and statistical techniques for high frequency phrase to process phrase string excavation, then take the basis of weight and weight fluctuations to compute hot-word weight. In the hot-word association period, hot words are divided based on the difference of the weight value of them and hot-word relationship was computed from the principle of co-occurrence rate (Li and Sun, 2011).

These methods provide good reference for us. However, the number of Tibetan webs is fewer than Chinese and the features of Tibetan information are different from Chinese. Therefore, some methods can not fit to Tibetan hot word detection.

In this study, we propose a comprehensive approach to Tibetan hot word extraction. Firstly, we use information entropy and vector space module similarity calculation to extract/filter Tibetan new hot words. Secondly, we use a frequency-position weighing algorithm to compute weight of word and combining entropy, 3σ criterion and variance to extract and track the Tibetan hot words. Finally, we construct word co-occurrence model to extract and analysis hot related words.

The rest of this study is organized as follows. The next section analyses the features of hot words. The extraction model of Tibetan new word is presented and discussed in section 3. Section 4 shows hot word extraction algorithm. And some experimental results by simulation are presented in section 5. Finally, the conclusion is drawn in section 6.

HOT WORD ANALYSIS

Typically, the hot topic word is mainly expressed in the following forms:

- Hot information is closely related to an important new event, so a lot of hot words are new words which have not been included in the dictionary. For example, "དམིགས་བསལ་ (SARS)" is a new word and hot word in a period of time. But it is usually segmented to "དམིགས་བསལ་" using word segmentation tools. In this case, the new valid hot words are difficult to be detected
- Hot words usually have characteristics of frequent occurrence, wide distribution and sudden transaction may occur over time. Some word appears with high frequency but little fluctuation which belongs to high-frequency words. Some word suddenly emerges transaction growth in one moment and has a rapid growth trend which belongs to hot word. Some word appears with low frequency and small fluctuation changes which belongs to low-frequency word
- Mutual information between related words more accurately reflects hot topic. For example, "ཡུལ་ཤུལ་/ས་འགུལ། (Yushu earthquake)" than "ས་འགུལ། (earthquake)" more accurately reflect hot topic. If only comparing the word frequency, the frequency of ས་འགུལ། will be higher. Therefore, degree of linkages between these words usually helps to bring out the hot events.

In this study, we will extract hot words and related words, shown in Fig. 1:

- Using Tibetan word segmentation and removing stop words to get Tibetan words
- Using information entropy and vector space module similarity calculation to extract/filter Tibetan new hot words
- Using word frequency position weighting algorithm to get the hot word candidate
- Based on entropy, 3σ guideline and variance statistical method, we get the hot words
- Using co-occurrence model to extract hot related words

TIBETAN NEW WORD EXTRATION

Typically, hot information is closely related to an important event, so some new words are usually hot words. So how to extract the valid strings from the segmentation fragments is the key technique of our research. In the study of Tibetan automatic word segmentation, we basically solved recognition of main Tibetan named entries using the "dictionary + rule + method.

So, we can use the following method to extract the new words:

- Using information entropy and uncertainty strategy to extract Tibetan new words

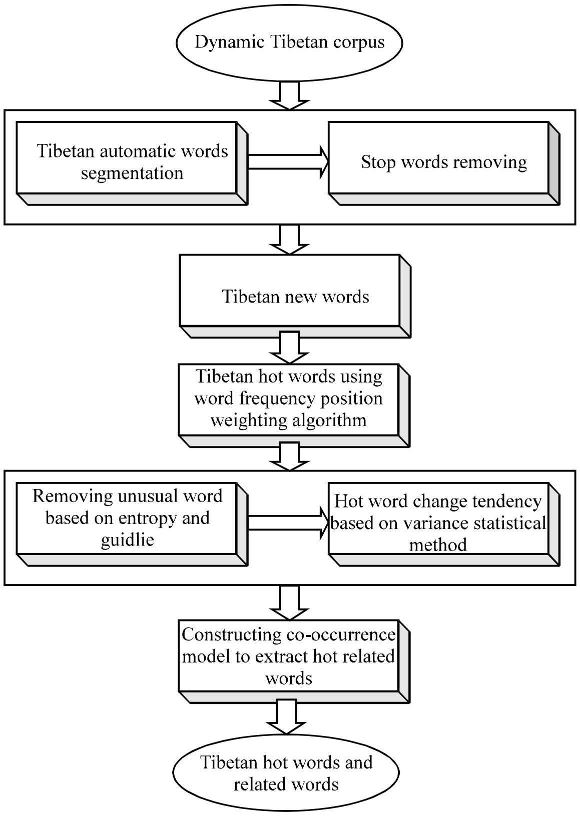

Fig. 1: Tibetan hot word extraction method

- Based on new word knowledge base, using vector space module similarity calculation to filter Tibetan new words
- Statistics and analysis frequency, distribution and usage of valid strings

Construction of electronic dictionary: The construction of Tibetan electronic dictionaries is the first step to corpus handling and processing. In this study, we construct the following Tibetan electronic dictionary:

- Tibetan basic word dictionary: we select "Tibetan and Chinese Dictionary" (upper and lower) edited by Zhang Yisu which is a comprehensive dictionary, collects more than 53,000 words
- Tibetan special words which are separated by two kinds: one is case-auxiliary word, the other is Tibetan named entity which is used to identify these special words. We collect Tibetan names from "Common Tibetan Name and Palace Name Dictionary", edited by Chen Guansheng and published in Foreign Languages Press, 2004. This dictionary contains 10,470 personal names and place names which can be used for Tibetan name recognition. Moreover, we collect 3,119 names from the corpus
- We have established Tibetan new word information electronic dictionary which contains 13,760 entries.

In establishment of Tibetan new word information electronic dictionary, we take "modern Chinese grammar information dictionary" as a model which constructed by Peking University Institute of Computational Linguistics. According to the "Modern Tibetan part of speech and tag set specifications for information processing", using classification and attribute description method, we describe grammar and semantic attributes of each word in detail. Major properties include:

- Conventional information (including word entries, meanings, syllable, e.g.)
- Grammatical information (including part of speech information)
- Word category information
- Word formation information
- Source of word

Using statistical method to establish Tibetan new word knowledge base: Through analysis of Tibetan new words, we find that the main ways of Tibetan new words are derivation words and compound words which use an inherent word as a morpheme.

- **Compound word:** Compound words are the main parts of Tibetan new words. There are six basic forms in Tibetan compound words: N+N, V+V, N+V, N+A, A+V, A+A. Using an inherent word as a morpheme to structure a new word which has new meaning. Such as འདྲ་ (net)+ཁང་། (house)→འདྲ་ཁང་། (internet), གློག (telegraph)+འཕྲིན། (information) →གློག་འཕྲིན། (telecommunications)
- **Derivation word:** Tibetan derivation words are mainly made up of suffix, infix and prefix. Tibetan language has ten traditional grammar typical suffix **(**པ་ཕོ་བ་བོ་མ་མོ་ཅན་ལྡན།**)**, infix **(མ)** and prefix **(ཨ).** Most of the basic vocabularies are derived. Such as ལས་ཀ་ (work)+སྒྲག་ (and other)+མི་སྣ་ (suffix)→ ལས་ཀ་སྒྲག་མི་སྣ་ (unemployed), སྲིད་གཞུང་ (government)+ལས་ཀ་ (work)+པ་ (suffix)→ གཞུང་དོན་པ (civil servant)

Based on Tibetan new word information electronic dictionary, we statistic character, syllable, word formation rules and word formation component frequency occurrence in the information electronic dictionary. Finally, we establish word morpheme attribute base of compound words and derivation words, get high form ability roots and affixes and establish Tibetan new word knowledge base.

Using information entropy and uncertainty strategy to extract Tibetan new words:

- Get all combination of segmentation fragments using total segmentation method
- Statistics frequency of combination blocks, get high frequency strings. Using information entropy and uncertainty strategy to extract Tibetan new valid words, get S_1

For high frequency string $c_1 \ldots\ldots c_n$, set threshold value a and b (a>b), if meet the following conditions, $c_1 \ldots\ldots c_n$ is a word.

- Frequency: $freq(c_1 \ldots\ldots c_n)>a$
- Left entropy: $H(C_0|c_1 \ldots\ldots c_n)>b$
- Right entropy: $H(C_{n+1}|c_1 \ldots\ldots c_n)>b$

Using vector space module similarity calculation to filter Tibetan new words: According to the Tibetan new words electronic dictionary and attribute base of compound word and derived word, filter S_1.

We establish M typical reference point using morpheme items in the attribute space. Through constructing morpheme compound words and vector space model (VSM), we extract the word with high similarity to the new word dictionary, initially identified as a legal word S_2.

Statistics and analysis frequency, distribution and usage of valid strings: Statistics and analysis on frequency, distribution rate and usage rate of valid strings S_2, obtain the candidate new word S_3.

Word distribution rate is:

$$D_i\ (\%) = \frac{t_i}{T} \times 100$$

where, t_i is text number of investigated word i appears, T is the total text number, D_i is word distribution rate.

Word usage rate is:

$$U_i = F_i \times D_i$$

where, F_i is frequency of word i, D_i is the distribution rate, U_i is the usage rate.

HOT WORD EXTRACTION ALGORITHM

Hot words usually have time highlights. It is not often mentioned in the past, but appears in media reports frequently at present. For example, "སྡུག་བསང (moral)" has been in the past, but with the immoral events appear, "སྡུག་བསང" has become the hot word.

Through text preprocessing, word segmentation, stop words removing and new word extraction in section III, we initially got a large number of words. Next, we use word frequency position weighting algorithm based on different importance of hot words in the title and content and Tibetan hot word extraction algorithm integrated with entropy, 3σ guideline and variance and finally get the hot words.

Word frequency position weighting algorithm: After pre-processing, we extract the contents of the Tibetan XML file in the title and content. Word Frequency Position Weighting algorithm (WFPW) is based on the position of the words, computing weight of word using Eq. 1:

$$W(x) = \begin{cases} T^2(x) \times \sqrt{C(x)} & T(x) \neq 0,\ C(x) \neq 0 \\ T^2(x) & T(x) \neq 0,\ C(x) = 0 \\ C(x) & T(x) = 0,\ C(x) \neq 0 \end{cases} \quad (1)$$

where, W(x) is the weight of word x, T(x) is the frequency of word x in the title and C(x) is the frequency of word x in the content. There are three cases:

- $T(x) \neq 0$, $C(x) \neq 0$ refers to the word x exits in the title and content
- $T(x) \neq 0$, $C(x) \neq 0$ refers to the word x exits in the title and does not exist in the content
- $T(x) = 0$, $C(x) \neq 0$ refers to the word x exits in the content and does not exist in the title

Removing unusual word based on entropy and 3σ guideline: We use the hot word entropy calculating to evaluate word stability. Entropy is too big, word stability is smaller. Entropy is too small, the meaning of the word is too low. So, we call word with higher entropy or low entropy as "unusual word".

First, we define the probability of occurrence P(x) and the entropy H(x) of the word x as following:

$$P(x) = \frac{n}{N},\ H(x) = -P(x)\log_2 P(x)$$

where, n is occurrence number of word x and N is the total number of words.

Second, we use 3σ guideline to further remove the unusual word, shown in Eq. 2:

$$P(\mu-\sigma<x\le\mu+\sigma)=68.3\%$$
$$P(\mu-2\sigma<x\le\mu+2\sigma)=95.4\% \quad (2)$$
$$P(\mu-3\sigma<x\le\mu+3\sigma)=99.7\%$$

where, σ is entropy standard deviation, μ is entropy average value x = μ. If the entropy of word x is not in (μ-3σ, μ+3σ), the word x is an unusual word.

Hot word change tendency based on variance statistical method: We use the variance-based method to calculate the change tendency of a word in a certain period of time, shown in Eq. 3:

$$S^2=\frac{1}{n}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right] \quad (3)$$

where, S^2 is variance of word x, n is the time length, x_1, x_2,…, x_n is the frequency of word x at each time point, $\bar{x}$ is the frequency average value of word x in one time period. The greater the variance which means the greater change degree of the word in this time and people give more attention to this word.

Constructing co-occurrence model to extract and analysis hot related words: Mutual information between related words can be more precisely covering the hot topic of the internet public opinion. With the development of a hot event, a group of hot words appear. We combine statistical method and co-occurrence word technology to achieve hot related words.

For the paragraph including the hot word w_i, we construct the co-occurrence model to statistic and extract the co-occurrence words w_m,…, w_n of w_i. We use the mutual information of words to evaluate the relevance of words, shown in Eq. 4:

$$MI(w_i,w_j)=\log\left[\frac{f(w_i,w_j)}{f(w_i)\times f(w_j)}\right] \quad (4)$$

where, $Mi(w_i, w_j)$ is the mutual information of word w_i and w_j, $f(w_i, w_j)$ is the frequency of word w_i and w_j in the same paragraph, $f(w_i)$ is the frequency of word w_i and $f(w_j)$ is the frequency of word w_j in test corpus.

Through computing $MI(w_i, w_j)$, we can get the co-occurrence words of the hot word w_i.

EXPERIMENTAL RESULTS

We statistic data of seven Tibetan webs from January to October in 2012 which contains 13,657 pages, 1,586,070 words.

After text preprocessing, word segmentation, stop words removing and new word extraction, we use word frequency position weighting algorithm to statistic the data change tendency from January to October.

Figure 2 and 3 show the difference of the word frequency change tendency in January. It is clearly that word frequency change is disorganized before using word frequency weighting position algorithm and the word frequency change curve is more obvious after using the algorithm.

Table 1 shows the unusual word after using entropy and 3σ guideline. The entropy of some word is too high or

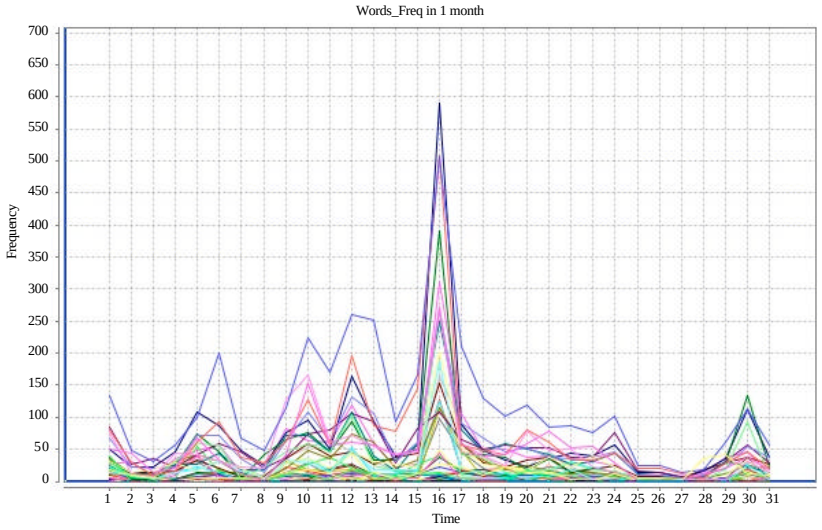

Fig. 2: Word frequency change in January not using WFPW

Table 1: Part of unusual words

Tibetan	Meaning	Entropy	3σ guideline
སྤྱིན་པ	Give	0.836	Unusual word
བྱེད་པ་	Do	0.179	Unusual word
འདི་	This	0.158	Unusual word
ཡོད་དའེ་	Have	0.109	Unusual word
ལན་	Answer	0.082	Unusual word
ཕྱིན་འཁོར་	Return home	0.547	Not unusual word
དགྲུན་ཚེ	Spring festival	0.522	Not unusual word
རིག་གནས་	Cultural	0.513	Not unusual word
དངུལ་འབྱོར་	Economy	0.508	Not unusual word

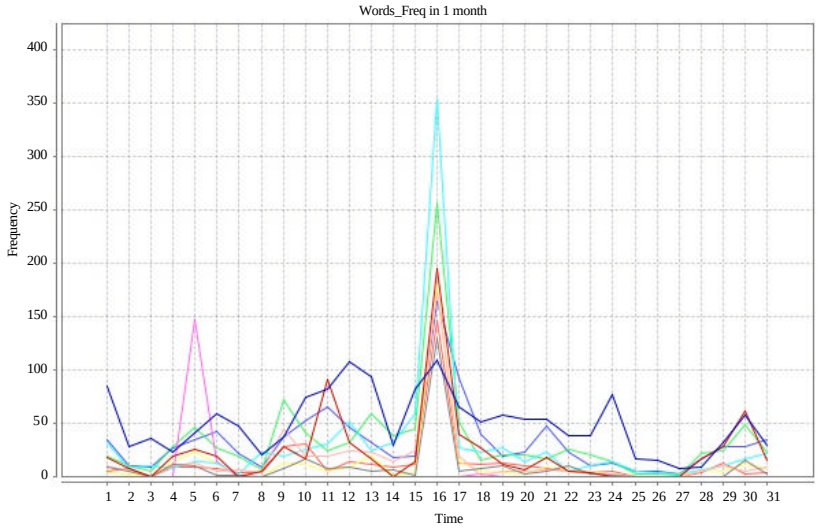

Fig. 3: Word frequency change in January using WFPW

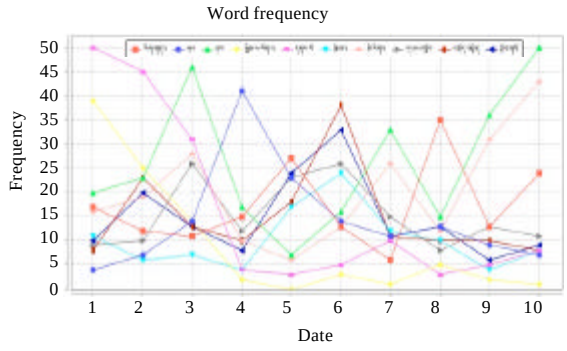

Fig. 4: Frequency change curve of the top 10 hot words

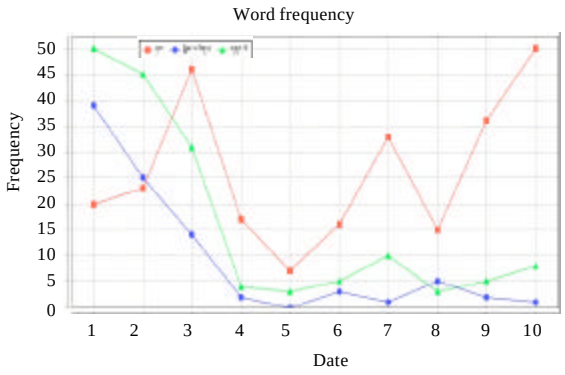

Fig. 5: Frequency change curve of some hot words

too low, such as སྤྱིན་པ (give), ལན་ (answer) and these words are unusual words. We will remove these words.

Finally, we get the top 10 hot words and the frequency change curve from January to October, shown in Fig. 4.

Figure 5 shows that hot word like དགྲུན་ཚེ (the Spring Festival), ཕྱིན་འཁོར་ (return home) which reflects people hope the home's living conditions in January and February during the Spring Festival. And the frequency of these words reduced from March and maintained a relatively low level from April to October. Meanwhile, the hot word དཔང་ (The Party) suddenly emerges transaction growth in March, July and October which reflects the related conference of the Party.

Meanwhile, we use co-occurrence model to extract the co-occurrence related words of top 10 hot

Table 2: Co-occurrence related words of top 10 hot words

Tibetan hot word	Meaning	Co-occurrence related words
ཀྱིམ་ལ་ཚོགས་པ་	Cultural	རྒྱང་ (Tibet), འཕེལ་རྒྱས་ (development), འཛུགས་ཐུབ་ (build), སྡེ་ཚོགས་ (society)
དཔྱིད་ཚེ	Spring festival	མཚོ་སྔོན་ (Qinghai), ཀྱིམ་ལ་ཚོགས་པ་ (return home)
ཀྱིམ་ལ་ཚོགས་པ་	Return home	མཚོ་སྔོན་ (Qinghai), དཔྱིད་ཚེ (Spring festival), སེམས་པ་ (mode)
སློབ་གསོ་	Education	སློབ་སྦྱོང་ (learning), མཚོ་སྔོན་ (course), མི་རིགས་ (Ethnic)
འཚོ་སྐྱོང་	Health	སྡེ་ཚོགས་ (safety), མི་རིགས་ (Ethnic)
ཀང་	Party	དམངས་འཐུས་ (people's deputies), ཞུ་ཚོགས་ (Committee), ཚོགས་འདུ་ (conference)
རིགས་གཅན་	Ethnic	ག་ངས་ཀྱང་ (minority), དམངས་ (people), མཐུན་སྒྲིལ་ (unite)
ཁྲིམས་	Law	འགན་འཁག་ (obligation), དབང་ (right), སློབ་སྦྱོང་ (learning)
ཁྱིམ་	Home	འཛུགས་སྐྲུན་ (build), ཁྲིམས་ (well)
དངུལ་འཚོང་	Economy	འཛུགས་སྐྲུན་ (build), རྒྱ་གཞུང་ (government)

words, shown in Table 2. From these words, we can get the hot topics of the internet public opinion more precisely.

CONCLUSION

In this study, we propose a model to automatically extract Tibetan hot words. We use information entropy and vector space module similarity calculation to extract/filter Tibetan new valid hot words. Based on the word frequency position weighting algorithm, we get the hot words. Finally, we construct word co-occurrence model to extract and analysis hot related words. Through analysis the dynamic Tibetan web data in 2012, the experimental results prove the model is effective.

ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation (No. 61331013), State Ethnic Affairs Commission Project (No. 12ZYZ010), National Language Committee Project (No. MZ115-94), National Social Science Fund (No. 11CY016) and the Fundamental Research Funds for the Central Universities (No. 1112KYZD05).

REFERENCES

Cai, Z.J., 2009. Blend in recognition of the Tibetan word segmentation system. J. Chin. Inf. Process., 23: 35-37.

Chen, L.Z., B. Li and X.P. Chen, 2012. Design and realization of monitoring system of campus BBS public opinion. J. Microprocessors, 2: 40-48.

Dai, Q.X., 2010. Information Processing Technology Research and Development of Minority Language. 1st Edn., Minzu Press, Beijing, China.

El-Beltagy, S.R. and A. Rafea, 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. J. Inf. Sys., 34: 132-144.

Jiang, D., 2006. Modern Tibetan verb syntactic and semantic classification and related syntax of sentences. J. Chin. Inf. Process., 20: 37-43.

Li, S.J., 2004. Research on maximum entropy model for keyword indexing. J. Comput., 27: 1192-1197.

Li, Y.Q. and L.H. Sun, 2011. Hot-word detection for internet public sentiment. J. Chin. Inf. Process., 25: 48-59.

Sun, Y., X.D. Yan and X.B. Zhao, 2009. Design of Tibetan word segmentation scheme. Proceedings of the International Conference on Information Engineering and Computer Science, December 19-20, 2009, China, pp: 1-6.

Sun, Y., X.D. Yan, X.B. Zhao and G. Yang, 2010. Research on automatic recognition of Tibetan personal names based on multi-features. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, August 21-23, 2010, Beijing, China, pp: 1-5.

Tomokiyo, T. and M. Hurst, 2003. A language model approach to keyphrase extraction. Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Vol. 18, July 12, 2003, Stroudsburg, PA., USA., pp: 33-40.

Wang, M., B. Li and C.Q. Sun, 2010. Research of network public opinion hotspots detection based on frequent items mining. J. Microcomput. Inf., 26: 35-38.

Xue, F., Y.D. Zhou and F. Gao, 2011. An online detection and tracking method for bursty topics. J. Xi'an Jiaotong Univ., 45: 64-69.

Yang, J., 2008. Keyword extraction in multi-document based on joint weight. J. Chin. Inf. Process., 22: 75-79.

Yang, Z., L.J. Duan and Y.X. Lai, 2010. Based on a short text string similarity clustering network. J. B. Univ. Technol., 36: 669-673.

Zheng, J.H. and J.L. Lu, 2005. Study of an improved keywords distillation method. J. Comput. Eng., 31: 194-196.

Zheng, K., X.M. Shu and H.Y. Yuan, 2012. Automatic discovery of network public opinion hotspot information. J. Comput. Eng., 36: 4-6.

Zheng, W., Y. Zhang, B.W. Zou, Y. Hong and T. Liu, 2007. Research of Chinese topic tracking based on relevance model. Proceedings of the Ninth National Conference on Computational Linguistics, April 10, 2007, China, pp: 58-63.