Hot Topic Extraction and Public Opinion Classification of Tibetan Texts

Guixian XU[1]*, Lirong QIU[1], Xueping PENG[2]
[1]Information Engineering College
Minzu University of China, Beijing, 100081, China
[2]The Centre for Quantum Computation & Intelligent Systems
University of Technolgy, Sydney, Australia
xuguixian2000@bit.edu.cn

Journal of Digital Information Management

ABSTRACT: *The increasing amount of Tibetan information has made Tibetan text processing popular and highly significant. In this study, Tibetan hot topic extraction and public opinion classification were investigated to accelerate the development of Tibetan information processing. First, Tibetan word segmentation in Tibetan hot topic extraction was presented. Second, feature selection based on term frequency and that based on document frequency was adopted to decrease feature dimensions. Third, a vector space model was used to conduct text representation. Finally, a statistical-based method was utilized to extract hot topics. In studying public opinion classification, a keyword table of public opinion needed to be established to conduct Tibetan public opinion classification. According to field, 18 classes were selected and used for public opinion classification. A keyword table of public opinion was constructed by domain experts. The approach to public opinion classification was introduced on the basis of the proposed similarity computation method. Depending on the proposed approaches, the application system was developed and used to carry out the experiments. Experiments show that the proposed method can extract topics effectively and classify public opinion rapidly. This research is helpful and meaningful for text classification, information retrieval, and construction of high-quality corpus.*

Subject Categories and Descriptors
I.2.7 [Natural Language Processing]: Text Analysis; **1.5.4 [Pattern Recognition]:** Applications – Information Extraction

General Terms: Data Mining, Information Extraction, Knowledge Management

Keywords: Hot topic discovery, Feature selection, Public opinion classification

Received: 17 January 2016, Revised 2 March 2016, Accepted 8 March 2016

1. Introduction

The rapid development of Tibetan network technology has allowed the use of forums and blogs to express one's opinion. Publishers often do not think about the authenticity and social effect of the content of their published works. A hot issue that spreads across networks immediately causes public concern and thus elicits related responses from people. In some instances, comments from the public can lead to considerable public pressure. The Internet, as a new medium, has become the platform for public ideological convergence and information dissemination. The study of hot topic detection in networks and public opinion classification is important to establish a harmonious environment that is conducive to public opinion expression and to maintain social stability.

Although Tibetan information technology has obtained several achievements, it still falls behind English and Chinese information technology. Important Tibetan information is buried in a large amount of Tibetan network data. Thus, effectively detecting hot topics online is becoming meaningful, especially because it can help monitor public opinion. Most studies have focused on topic detection and tracking (TDT) technology [1]. The TDT technique can effectively collect and organize scattered information about an event. A topic generally represents the topic of the documents of a class. Several studies have

established topics via the automatic clustering of large-scale documents [2]. People prefer a simple topic. Accordingly, topic representation has been studied [3, 4]. Multi-document summaries based on certain topics [5] have also been proposed. Unlike the TDT clustering algorithm based on a large number of documents, the keyword topic detection algorithm can extract hot topics from documents with similar contents. As a result, users can browse and focus on recent and important event topics.

Document classification has been widely studied since the early 1960s. The earliest method for document classification is the word matching method. This method determines whether a document belongs to a category on the basis of whether the document contains words that are the same as the name of the category. Later on, the statistical learning method was developed to classify texts. This method is known for its high classification accuracy. Many classification technologies have theoretical bases and clear evaluation criteria. Thus, the statistical learning method has become the mainstream approach in the text categorization field.

The present study is a technology research into Tibetan hot topic extraction and public opinion classification. The remainder of this paper is organized as follows. Section 2 describes the literature review. Section 3 introduces the methodology. Section 4 provides an analysis of the experimental results. Section 5 summarizes the conclusions.

2. Literature Review

The study of hot topic extraction technology began in the mid-1990s. Studies at that time mainly focused on association detection, topic detection, subject tracking, and cross-language TDT [6, 7]. The analysis of network public opinion has achieved some success, but many problems remain unresolved. In several existing network public opinion monitoring systems, the keywords used in the monitoring process is defined by certain individuals. Owing to the limitations related to the knowledge of these individuals, as well as the limitations to information source and many other subjective factors, hot topics pertaining to emergencies cannot be monitored. However, a hot topic discovery system can be developed through information technology. Such a system can organize and analyze emergency news and automatically identify hot event topics. The system can also handle the real-time tracking of unexpected events.

Christian [8] identified the statistical distribution characteristics of feature words in a text corpus and measured the similarity of the words. The k-means algorithm was used to cluster the texts for topic detection.

James Allan [6] and Yiming Yang [7] established online identification systems to inspect emerging events. Zhao [9] proposed a topic-oriented community detection approach that combines social objects clustering and link analysis. A subspace clustering algorithm was used to group all the social objects into topics. Then, the members involved in those social objects were divided into topical clusters. In differentiating the strengths of connections, a link analysis on each topical cluster was performed to detect the topical communities. Experiments showed that the approach was able to identify more meaningful communities. The research on structured topic models was first conducted by Zhao Hua [10] and Jin Zhu [11]. The former conducted timing and distribution density identification of topic evolution on the basis of boundaries. The latter clustered relevant reports, established different feature extraction clustering techniques in an event frame to describe topics, and formed events within a framework of modality relationships and roles through HowNet to help describe the tendency of different aspects of a subject. The two methods can improve the performance of topic tracking systems.

Common automatic text classification algorithms fall into two categories. The first category comprises the classification algorithms based on probability and information theory, such as the naive Bayes method and the maximum entropy algorithm [12]. The second category comprises the classification algorithms based on knowledge study, such as decision trees [13], artificial neural networks [14, 15], and support vector machines [16].

Ontology is a clear formal specification of the shared conceptual model [17, 18]. This specification has the characteristics of conceptualization, explicitness, formalization, and sharing. Thus, ontology construction can achieve some degree of knowledge sharing and reuse. Existing classification methods are based on ontology, and training samples can be obtained through ontology semantic information and the text categorization implementation of automatic text classification. Semantic classification has obviously become a hot research topic in the field of text classification [19, 20].

3. Methodology

The proposed method is introduced in this section. First, hot topic extraction is discussed. Second, the approach to public opinion classification is presented.

3.1. Hot Topic Extraction

The construction of the dataset is described, and Tibetan word segmentation is presented. Feature selection and text representation are conducted. The classical term frequency–inverse document frequency (TFIDF) [21] is used to calculate the weights of features. A statistical-based method is utilized to extract hot topics.

3.1.1. Dataset Construction

A web spider is used to download web pages from http://www.qhtb.cn/. Considering that the content of a HyperText Markup Language file is irregular and has several different formats, we use the rules of regular expressions to extract important information from the web pages, including

dates, text titles, text subtitles, author names, and Tibetan texts. Then the web pages are deleted with redundant information. Finally, we utilize an extensive markup language (XML) file to store the information from the web pages.

3.1.2. Word Segmentation

Given that text information is unstructured, computers cannot understand it. Word segmentation is indispensable in machine learning. This process splits texts into words that computers can store and handle before information processing. Word segmentation is easy in English because spaces and punctuations can be utilized to split text information. By contrast, Tibetan word segmentation is a difficult process and a great challenge because the Tibetan words do not include natural space separators. Syllables serve as Tibetan word units. A Tibetan sentence is composed of syllables. Tibetan electric dictionary, place name dictionary, person name dictionary, and case auxiliary words are used to split a Tibetan text. The word segmentation of the Tibetan text in this study is shown in Figure 1.

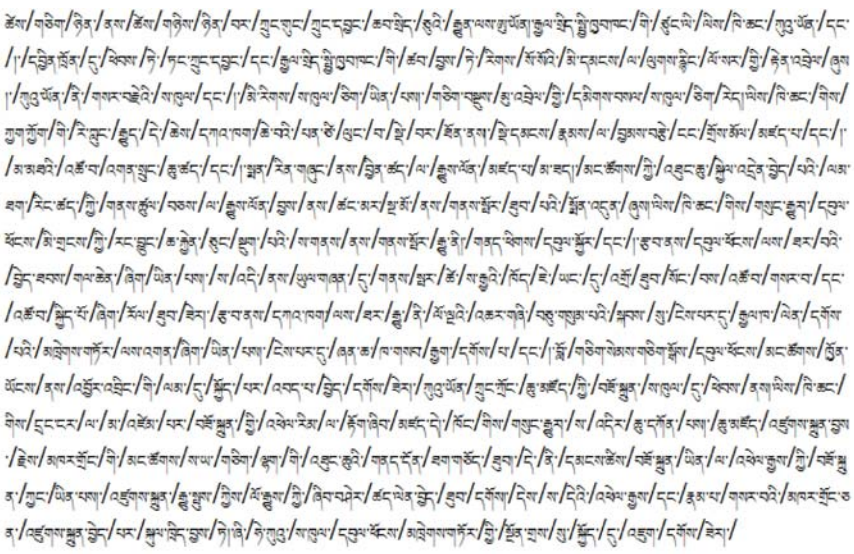

Figure 1. Word segmentation of a Tibetan text

3.1.3. Feature Selection

During document feature extraction, the dimensionality of the original feature space is usually high. A large number of irrelevant or redundant features are produced [22, 23]. For example, a training set with approximately 1,000 texts may generate thousands of feature items. Feature dimensionality must thus be reduced to decrease learning size and to solve the dimensionality problem. The features within a feature carry sufficient information to express texts. Feature selection refers to the process of selecting a feature subset from original features that best reflects the statistical characteristics of a text.

We use the term frequency (TF) and document frequency (DF) methods as feature selection approaches to reduce the dimension of a feature space and to improve the efficiency and accuracy of the topic extraction process.

(1) TF

TF refers to the number of feature sets that appear in a training text. TF assumes that a term with a frequency greater than a threshold contributes greatly to information processing. Thus, terms of high frequency are retained in the original feature set by setting a threshold value. Terms with a frequency lower than the threshold are filtered out.

(2) DF

DF of a term refers to the number of documents in which the term appears. The use of DF as a feature selection method is based on the following assumptions: if the DF value of a term is below a certain threshold, then its contribution to topic extraction is not significant. Hence, the term is removed from the original feature space. However, if the DF value is too high, e.g., a certain term emerges in each document, then its contribution to topic extraction is also not significant. Generally, features of low DF value are filtered out, whereas features of high DF value are disregarded in the original feature space to allow the formation of a low-dimensional feature space.

3.1.4. Text Representation

The vector space model (VSM) [24] uses a bag of words; a text is treated as an unordered set of features. In the VSM, $D = \{d_1, d_2, \ldots d_m\}$ is a collection of documents. The feature vector set of a document set $T = \{t_1, t_2, \ldots t_n\}$ is generated by feature selection methods. N is the total number of features, i.e., the dimension of the feature space. Each document d is represented as a vector $v(d)$, $v(d) = \{w_1, w_2, ..., w_n\}$. The i th component of vector w_i is the weight of feature t_i. $TFIDF$ is used as the term weighting method. For candidate keywords, we calculate the TF value of each word in the text and the inverse DF(IDF) value of each word in the document set. A term with a high TFmeans that it has a high degree of concern of the topics in the text. A term with a high IDF means that it expresses the meaning of this text to a large degree. The

formula is shown [23] as follows.

$$TFIDF(t,d) = TF * IDF \qquad (1)$$

$IDF = \log(|D|/|DF|)$, where $|D|$ represents the total number of documents in a training set.

3.1.5. Hot topic Extraction

Each hot topic can be utilized to identify the original sources of information. Topic results can be obtained according to different modes, e.g., (1) extracting hot topic words from an XML file set, (2) extracting hot topic words from an XML file set depending on the data scope, and (3) extracting hot topic words from an XML file set depending on the time interval.

The detailed extraction algorithm is as follows:

Input: The number of hot words n.
A Tibetan XML file set.

Begin:

(1) For every XML file, execute word segmentation.

(2) Conduct feature selection, and form a feature space.

(3) Calculate TFIDF value w_i of every term t_i of each file that belongs to the feature space.

(4) Compute the sum of every w_i of all texts where t_i appears. The sum of is w_i called S_i.

(5) Sort S_i of every term t_i in descending order.

(6) Select n terms of n highest S_i as hot topics.

(7) Output n hot topic words.

End

3.2. Public Opinion Classification Approach

3.2.1. Construction of Keyword Table for Public Opinion Classification

A keyword table of public opinion needs to be established to conduct Tibetan public opinion classification. According to field, 18 classes are selected and used for public opinion classification. These classes include natural disasters, accidental disasters, public health, education reform, social security, anti-corruption, forced demolition, monopoly enterprises, social ideological trend, and supervision of public opinion. The number of keywords differs for every class. The detailed information is shown in Table 1.

Figure 2 shows examples of the keywords of the accident disaster class. The first column comprises the Tibetan words. The second column comprises the corresponding English meanings of the Tibetan words.

Class Name	Number of Keywords	Class Name	Number of Keywords	Class Name	Number of Keywords
Urban management	64	Anti-corruption	82	Public health events	17
Education reform	154	Monopoly enterprises	85	Compulsory removal	94
Anti-pornography	25	Social security	118	Social ideological trend	48
Accidental disasters	61	Network politics	118	Cultural disputes	92
Doctor–patient relationship	176	Public opinion supervision	113	Natural disasters	122
Total			1369		

Table 1. Keyword table of public opinion classes

3.2.2. Public Opinion Classification

Inorder to complete public opinion classification, keyword tables need to be constructed. The number of keyword tables is assumed to be n. $\{C_1, C_2, \ldots, C_n\}$ express n keyword tables. C_i comprises m keywords $\{t_1, t_2, \ldots, t_m\}$, $i = 1,2,...,n$. For each C_i, m is not equal. Assume d_x is the Tibetan text needed to be classified. We use d_x to match the keyword table C_i and d_x to determine the frequency of the appearance of the words in C_i. The similarity between C_i and d_x is then computed. The similarity formula is as follows:

$$sim(C_i, d_x) = 0.3 \times (s/m) + 0.4 \times r + 0.3 \times p \qquad (2)$$

where s is the number of similar words between C_i and d_x. m is the number of keywords of C_i. s/m is the proportion of d_x to C_i. If the value is high, then the possibility of d_x belonging to C_i is strong. r is the frequency sum of s words. p is the average word frequency, i.e., $p = r/s$. High r and

རླངས་འཁོར་མེ་སྐྱེན།	car fire accident
འཕྲིན་འགྲུལ་འཁྲ་ཀག་གི་དོན་ཆུབ།	railway traffic accident
དམངས་ཚེད་མཁའ་སྐྱོད་འཕྲུལ་གྱི་དོན་ཆུབ།	aircraft accident
སྐོག་སྐབས་དོན་ཆུབ།	elevator accident
སོལ་དུག་སོག་པ།	gas poisoning
ཉིང་རྡུལ་དོན་ཆུབ།	nuclear accident
འབབ་མའེལ་འཛདབཀོལ།	bomb attack
སྐོག་འབྱུང་པ།	electric shock
རྒྱ་དིམ།	drowning
མར་ཧགས་པ།	falling injury
ཕུག་གསས་གང་འཛགས།	personal security
བའེ་ཚང་གཙོད་འཛོ།	health damage

Figure 2. Keywords of the accident disaster class

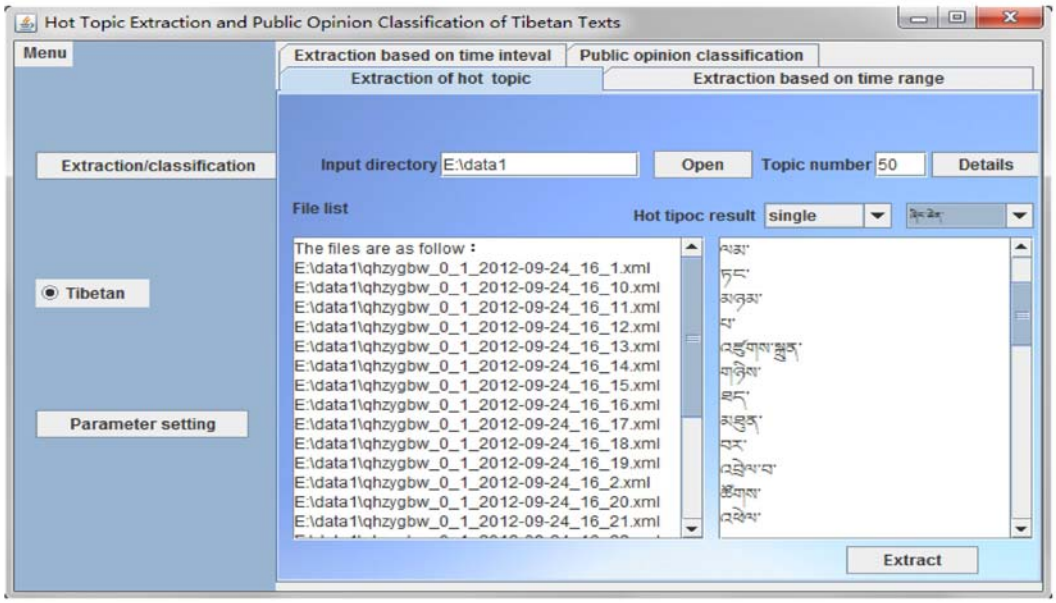

Figure 3. The result example of hot topic extracion

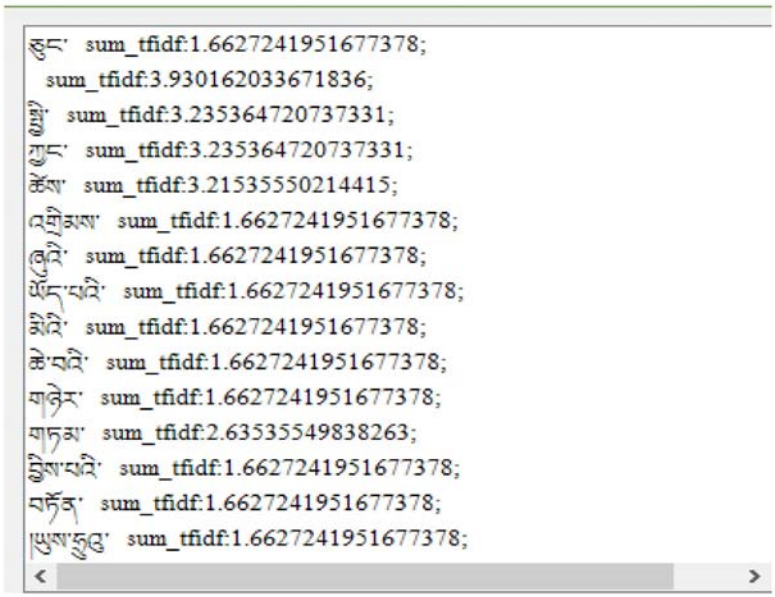

Figure 4. Graph of the sum of the TFIDF of each word

p values indicate a great likelihood for d_x to belong to C_i. 0.3, 0.4, and 0.3 denote the importance weights of s/m, r, and d, respectively.

$sim(C_i, d_x)$ is computed, and the maximum similarity value among all $sim(C_i, d_x)$ $(i = 1,2,...,n)$ is selected. d_x is then classified to this class.

4. Result Analysis and Discussion

We use XML texts to conduct the experiment in this work. For hot topic extraction, we set the number of hot topics to 10 and the number of files of the dataset to 50. We set TF to be equal to 3 and DF to be equal to 2 for feature selection. We execute the application system. The result example of hot topic extraction is shown in Figure 3. Figure 4 shows the sum of the TFIDF of some words.

We increase the size of the dataset and evaluate the TFIDF sum of three hot words (ཡོད་པའི་ , ཐོང་ , and རྫོད་པའི་). The increasing size of the dataset is limited to ensure that the three hot words appear in the hot topic results of every data set. Table 2 shows the TFIDF sum of the three words from Table 2. The TFIDF sum of each word increases with increasing dataset size. The TFIDF value of each word is determined by TF and DF. The weight calculation method of TFIDF can be used to effectively extract the hot topics from the experiment data.

Dataset size	TFIDF sum of ཡོད་པའི་	TFIDF sum of ཐོང་	TFIDF sum of རྫོད་པའི་
20	1.116962261202364	1.116962261493756	1.1169622619357354
50	1.2842188790692597	1.28421884753866	1.2842188936295525
100	1.6489253373124755	1.648925323743632	1.6489254793264337
150	1.770343298726456	1.770347922346924	1.7703433677835249

Table 2. TFIDF sum of three words in different dataset sizes

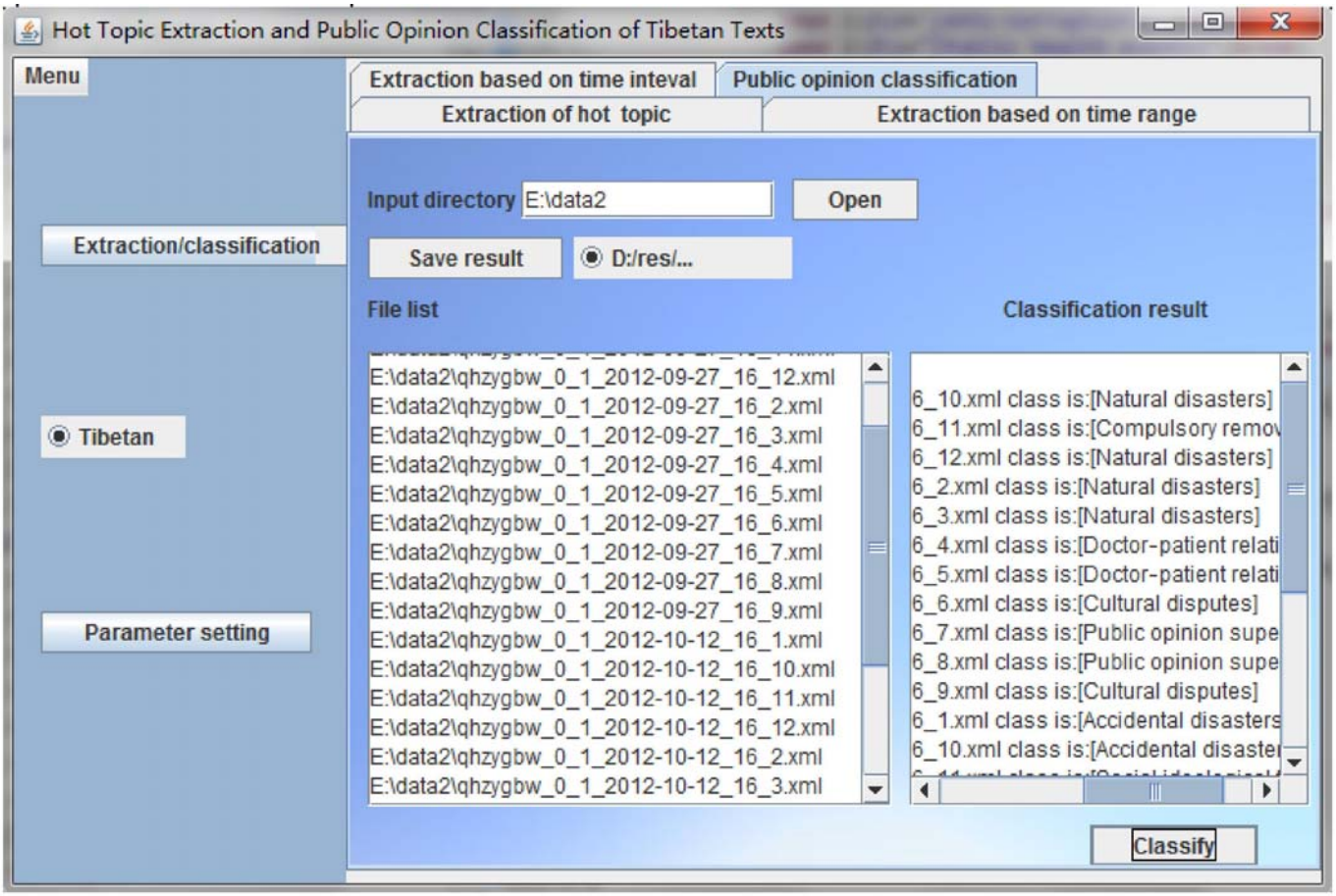

Figure 5. Interface of the public opinion classification tool

Figure 5 shows the interface of the public opinion classification software. We can select the directory, including the XML files needed to be classified. The classification result path can also be set. Figure 5 illustrates nine XML files that are classified according to the classes of cultural disputes, social security, and anti-corruption. The classification performance is 80%. The proposed classification method is effective for classifying public opinion. However, the classification performance is affected by the quantity and quality of the keyword table of Tibetan public opinion. If the quantity and quality of the keywords are improved, then the classification performance is enhanced further.

5. Conclusions

This paper presents a method for finding hot topics that is based on a statistical approach. The experiment results show that the application system can extract topics effectively and that the results can reflect the characteristics of hot topic categories. The approach to public opinion classification involves the use of a keyword table. This approach could rapidly classify Tibetan texts of public opinion. However, the classification performance is related to the quantity and quality of the keyword table. If keywords are improved, then the classification accuracy will be enhanced further.

The aim of hot topic extraction and the public opinion classification of Tibetan texts is to facilitate the way in which users classify information. This research can effectively contribute to the identification of public emergencies and is helpful and meaningful for text classification, information retrieval, and construction of high-quality corpus.

Acknowledgment

This work was supported by theNational Key Technology Research and Development Program of the Ministry of Science and Technology of China (No.2014BAK10B03), the Beijing Social Science Foundation (No. 14WYB040), and the National Natural Science Foundation of China (No. 61309012, No. 61331013).

References

[1] Zhang, XY., Wang T(2009). Topic discovery and tracking technology research. *Computer Science Exploration*, 3 (4) 347-357.

[2] Long, ZY., Cheng, W(2011). Kind of hot topic detection algorithm based on clustering key words. *Computer Engineering and Design*, 32 (6) 2214-2217.

[3] Hang, L., Kenji, Y(2003). Topic analysis using a finite mixture model. *Information Processing & Management*, 39 (4) 521-541.

[4] Pons-Porrata, A., Berlanga, L., Ruiz-Shulclper(2007). Topic discovery based on text mining techniques. *Journal of Information Processing and Management*, 43 (3) 752-7 68.

[5] Wan, XJ., Yang, JW.(2008). Multi-Document Summarization Using Cluster-Based Link analysis. *Proceedings of 31st International Special Interest Group on Information Retrieval conference*, Singapore: ACM, p. 299-306.

[6] Allan, J., Papka, R., Lavrenko, V(1998). On-Line New Event Detection and Tracking. *Proceedings of 21st International Special Interest Group on Information Retrieval conference*, Melbourme, Australia: ACM, p. 37-45.

[7] Yang, Y., Pierce, T., Carbonell, J(1998). A study on Retrospective and On-Line Event detection, *In*: Proceedings of 21st international Special Interest Group on Information Retrieval conference, Melbourme, Australia: ACM, p. 28-36.

[8] Wartena, C., Brussee, R(2008). Topic detection by clustering keywords. *Proceedings of 19th International Conference on Database and Expert Systems Application*, Turin, Italy: IEEE, p. 54-58.

[9] Zhao, ZY ., Feng, SZ., Wang, Q., Joshua, ZH., Graham, J., Williams, Fan JP(2012). Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, (26) 164-173.

[10] Zhao H, Zhao TJ, Yu H, Zhang S(2006). Dynamic evolvement-oriented topic detection research. *Chinese high technology letters*, 12 (16) 1230-1235.

[11] Jin Z, Lin HF, Zhao J(2005). Study on Topic Tracking and Tendency Classification Based on HowNet. *Journal of the China society for scientific and technical information*, 5 (24) 555-561.

[12] Kazama J, Tsujii J(2005). Maximum entropy models with inequality constraints: A case study on text categorization. *Machine Learning*, 60 (1) 159-194.

[13] Dewan MF, Li Z, Chowdhury MR, Hossaina MA., Rebecca S(2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41 (4) 1937-1946.

[14] Dêbska, B., Guzowska-Œwider B(2011). Application of artificial neural network in food classification. *Analytica Chimica Acta*, 705 (1) 283-291.

[15] Koushal k. Knowledge Extraction From Trained Neural Networks. *International Journal of Information & Network Security*, 1(4) 282-293.

[16] Ghamisi, P., Couceiro, MS (2014). A Novel Feature Selection Approach Based on FODPSO and SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (5) 2935-2947.

[17] Chen, EH., Wu, GF(2005). An Ontology Learning Method Enhanced by Frame Semantics. *In*: *Seventh IEEE International Symposium on Multimedia*, Irvine, Callifornia: IEEE, p, 374-382.

[18] Karyawati, A.A.I.N. E., Winarko. E., Azhari, A, Harjoko

A. (2015). Ontology-based Why-Question Analysis Using Lexico-Syntactic Patterns. *International Journal of Electrical and Computer Engineering*, 5 (2) 318-332.

[19] Huang XT. (2009). Research on semantic Web text classification based on Ontology. *Library*, 3 (3) 47-49.

[20] Tsytsarau, M ., Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24 (3) 478-514.

[21] Pedram VA, Omid SSh. (2015). Scientific Documents clustering based on Text Summarization. *International Journal of Electrical and Computer Engineering*, 5 (4) 782-787.

[22] Girish, C., Ferat S. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40 (1) 16-28.

[23] Min, F., Hu, QH., Zhu, W. (2014). Feature selection with test cost constraint. *International Journal of Approximate Reasoning*, 55 (1) 167-179.

[24] Katrin, E. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6 (10) 635-653.