A review of Technologies on Tracking Tibetan Public Opinion Topics

Bi-Rong CHEN[1], Gui-Xian XU[2,a]

School of Information Engineering,Minzu University of ChinaBeijing 100081,China
[1]913756830@qq.com,[2]xuguixian2000@bit.edu.cn

Abstract. The target of technologies on tracking Tibetan-based network public opinion topics is to track the development of some known topics. By tracking these topics could help people to grasp the trend of public opinion. This research area becomes more and more important on Natural Language Processing domain and how to apply these state-of-the-art computer science technologies on the study of national language has drawn lots of awareness to related researchers. This article will introduce some features about Tibetan language and its grammar. Some technologies about categorizing Tibetan text and tracking Tibetan-based network public opinion will also be talked about.

1 Introduction

With the rise of social media, more and more professional researchers pay attention to the public opinion generated from social applications, such as Weibo, Tieba. Public opinion indicates a reflection of individuals or communities' reaction on the social phenomenon which are related to their own concern or profit. During a period of time, public opinion could influence people's emotion and lead to conformity. So, it is very important to analysis public opinion and prevent some harmful public opinions from misleading people's emotion. In Tibet, with the booming of economy and support of central organization on development of science and technology, the number of netizens in Tibet grows significantly. Unverified information is exaggerated or distorted by the internet usually affects national unity and social stability. It is not reality to let humanbeings to filter public opinion, since we are in an information explosion age. That will cost lots of recourses and amount of time to handle that. So, applying start-of-the-art technologies on supervise public opinion of minority languages is a good approach and that will support the government on supervise public information.

2 The situation of tracking Tibetan-based network public opinion in China and abroad

2.1 Features of Tibetan language and grammar

[1]introduced Tibetan language belongs to phonetic system, it consists of consonants, vowels and punctuation symbols. There are thirty consonants.As shown in Figure 1.

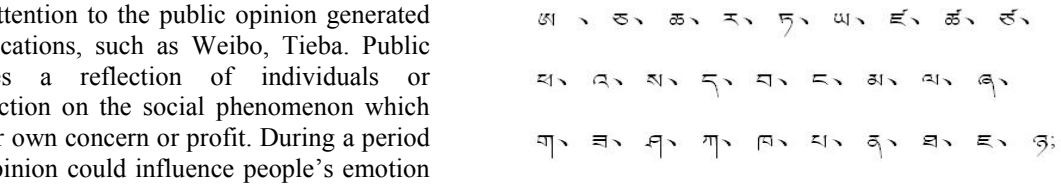

Figure 1. Thirty Consonants

There are four vowels, as shown in Figure 2.

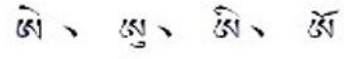

Figure 2. Four Vowels

According to the different position of syllable, Tibetan character could be divided into six categories, which are "base-word", "adding before word", "adding after word", "adding top of word", "adding bottom of word" and "double adding after word". [2]introduced in the article named The specification of Tibetan character category based on information processing. Tibetan character has been categorized into 29 base cluster and 6 special clusters.

2.2 Situation in abroad

The research of public opinion in western counties start more earlier and develop very fast. For example, [3] started the research on public opinion since 1973.[4] introduced a basic cluster algorithm called Ant colony algorithm. [5]applied this algorithm on text analysis.[6]

invented a system called Opinion Finder, which could automatically find the sentence high related to the subject or the terms in a sentence which are relate to the subject.

2.3 Domestic situation

The research on public opinion tracking begins a bit late in China, and some technologies specific on Tibetan language are also not perfect. But with the development of the technology.[7]applied a distributed solution of Multi- Agent algorithm on Tibetan text clustering.[8]combined entropy and conditional random field models to recognize Tibetan names and get a good performance, the accuracy is above 93%.[9]combined emotion pattern and machine learning algorithm and introduce a new algorithm called PMML.

3The technique detail of tracking network public opinion

3.1 Text pre-processing

3.1.1 Remove noise of whole Tibetan web page

Web pages contains two kinds of information, one is subject related information and the other one is off-subject information, which contains pictures, videos etc. we call all of these off-subject information as “web noise”. The appearance of web noise not only cause ambiguity of web subject but also influence the tracking of web subject information. So, before analysis the web page, all noises should be removed at first. A cleaner web text could significant increase the performance of categorization.

3.1.2 Tibetan text decode

[10]introduced Tibetan language belongs to phonetic system. The Tibetan National Standard formulated in 1997 has set some rules for Tibetan character encoding and a set of base Tibetan characters.Tibetan script in the memory of the computer requires a plurality of two-byte characters to represent the Tibetan script Ding.But Tibetan based on web coded in different ways, Have the same element code, Himalaya, basic set, spread code, Pandita,etc. [11]convert different coding Tibetan to Unicode format,in order to better analyze the Tibetan public opinion tracking and follow-up work.

3.1.3 Tibetan segmentation

Tibetan language is a phonetic language, which differs from English, no special symbol mark between the Tibetan word for segmentation.[12]The theme of representation and classification are based on good segmentation results,so merits of corpus participle will directly affect the accuracy of the Tibetan public opinion analysis.For general segmentation method has the following four: Mechanical word segmentation,Rule word segmentation,Statistics word segmentation,and Combine statistics and rules word segmentation.

3.1.4 Removing stop words

After the segmentation of the corpus, the corpus text will be a collection of words.Like many languages, Tibetan also has a lot of stop words.[13]believed that in addition to digital, daily life too common nouns, pronouns and verbs have no actual meaning can also be divided into within the stop word category.The general use of word list to remove the stop words.[14]introduced construct a stop word list procedure can be divided into two steps:first, the corpus of function words, pronouns, conjunctions, auxiliary and other stop words added into word list,then calculate the frequency of the word through the corpus, selected frequency is higher than a certain threshold, these words will also join in the word list.

3.2 Text representation

Text is often regarded as a string consisting of many characters, but this form is not conducive to the machine to do the classification and learning,So first we should translate the text to be processed into a format that is easy for machine learning and classification..Commonly used text representation methods include: vector space model (VSM), probability model, Boolean model, language model. For the main Tibetan network public opinion analysis model is vector space model and language model.

3.3 Text similarity comparison

[15]introduced text similarity is a statistic that used to measure the degree of similarity between two text, generally used to represent.

3.3.1.Inner product between vectors

The larger the value of the inner product obtained, the greater the similarity, the text D1, D2 vector inner product of similarity is calculated as follows:

$$Sim\ (D_1, D_2) = \sum_{k=1}^{n} W_{k1} \times W_{k2} \qquad (1)$$

3.3.2 Cosine similarity

This method control the value of the similarity between -1 to 1, the cosine is calculated news text D1 and D2 similarity such as:

$$Sim\ (D_1, D_2) = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} w_{1k}^{2}\right)\left(\sum_{k=1}^{n} w_{2k}^{2}\right)}} \qquad (2)$$

3.3.3 The maximum and minimum method

The maximum and minimum method, the calculation of this method is simple, also can through the mathematical calculation, remove maximum minimum brings difference in each category.

$$Sim\ (D_1, D_2) = \frac{\sum_{k=1}^{n} Min\ (W_{1k}, W_{2k})}{\sum_{k=1}^{n} Max\ (W_{1k}, W_{2k})} \quad (3)$$

3.4 Feature extraction and weight calculation

3.4.1 Feature extraction

After segmentation of the Tibetan text in the network, has a series of features set,Feature selection is to extract the effective feature set from the original feature set, remove the original feature set some weight over the small feature set, to achieve the purpose of dimension reduction,improved the efficiency of clustering , and the complexity of time and space is reduced.

Commonly used feature selection methods include: Document Frequency, Mutual Information, Information Gain, Estimation of, χ^2 etc.

3.4.2 Weight calculation

There are many methods to calculate the weight of feature items.Such as TF algorithm, IDF algorithm, tf*idf algorithm and mutual information algorithm, etc..[16]Among them, the tf*idf algorithm combined TF with the IDF algorithm is the most widely used algorithm.The formula used Tf*IDF algorithm calculate the weights of the feature item as shown in (4):

$$w_i = tf_i * idf_i * f_i(w) \quad (4)$$

Where tfi represents the frequency of a feature item i in the document,idfi represents inverted file,fi(w) indicates the weight given to the feature item of named entity or a in the title.

3.5 Public opinion tracking technology

[17]introduced topic tracking process is to detect new topics or follow-up related reports in the follow-up news.The basic process is , first to detect new news, and if it is not aware of the relevance of topics, it is considered a new topic, and if you can find it linked to the subject , it will be the news text classification among this topic.topics Tracking is different from the traditional classification algorithm, The traditional classification algorithms have identified the number of categories is the number of topics, but topics tracking technology is a dynamic classification techniques.Over time, constantly testing the new text,the number of topics will increase because detect the text does not belong to the pre-existing topics.

Tibetan network public opinion tracking technology in fact is equivalent to a mapping relationship .Mapped The hot topic which was later detected to the topic categories which were classified by the training data.Public opinion tracking technology includes two aspects, the first is the discovery of the topic, the second is the topic of tracking.

3.5.1 topic discovery technology

In this part , the main work includes:
a).Selection of feature items from training samples;
b).Representation the feature item we choose;
c).Through a number of classification algorithms to achieve our discovery result.

The choice of feature items and the representation of the feature items have been introduced respectively in3.2, 3.4, and the classification of the topic is the application of text classification technology.

The text classification of Technologies on Tracking Tibetan-Based Network Public Opinion Topics is an important branch of machine learning and data mining direction,so, many algorithms based on machine learning can be used for text classification.[18]introduced text classification algorithms can be roughly divided into two categories: rule-based algorithms and statistical-based methods,Rule-based algorithms include decision tree algorithm and Rough Set;the text classification based on Statistical algorithm can be divided into unsupervised learning, supervised learning and semi-supervised learning,including the k-nearest neighbor algorithm, least squares fitting,Neural Networks,Support Vector Machine, Rocchio and Naive Bayesian.At present, most classification algorithms are all supervised machine learning.It contains the training data and test data ,hrough a collection of training data is formed classifier,Then use the classification processing text to be analyzed,Therefore, use of which algorithm structure text categorization It is an important step in the ability to accurately track the Internet public opinion.

(1)Naive Bayesian

Theoretical Basis of Naive Bayesian is Bayes Theorem,In seeking the probability of an event provided another one occurred at a time

$$P(A \mid B) = \frac{P(B \mid A)P(B)}{P(B)} \quad (5)$$

[19]introduced use Bayesian probability algorithm applied to text categorization is:When a text appears, the probability of It belongs to each of Subject categories,Depending on the size of the calculated probability, can get the category the text belongs to.

[20]introduced the shortcomings of the naive Bayesian algorithm:

First, we use Bayesian formula for classification, assumption that the feature are independent of each other, each feature is not affected by other feature.But in real life, this assumption is difficult to set up.There is a mutual relationship between feature items, which

indicates that the feature items may not be completely independent.

Second, using naive Bayesian algorithm to train the classification samples, we need a lot of training data to get the probability of each feature item appears in each category,It will cost a lot of energy and time.

(2)KNN algorithm

KNN algorithm, also known as K-nearest neighbor algorithm.In 1967, proposed by Cover Hart, It is an inert algorithm , based on the instance.The basic idea of KNN Algorithm is set traversal similarity comparison to the newly added text with training text.selected K text of most similarity,According to the text of subject categories to determine the categories of newly added text.[21]

K-NN algorithm is relatively simple, its classification effect is better, but also has its own shortcomings:

The first one is to compare the ergodic similarity between the new text and the text of the training set,So you need to store all the training set of text, Storage capacity is very large.

Second, the classification effect of KNN algorithm depends on the choice of K value, how to choose the appropriate K value is very critical, which has become an urgent task of the KNN algorithm to achieve.

(3)Support Vector Machine(SVM)

Support Vector Machine seeking the optimal categories on the case of linear separable ,Wherein the optimal classification refers to the ability to conduct two classes of error-free division,and to ensure that the maximum interval between these two classes.Core content of SVM made by Vapnik between 1992 with 1995,this algorithm based on Structural Risk Minimization Theory and Statistical Learning VC Theory .

[22]introduced the basic idea of SVM classification algorithm is: First, the training text is represented by a vector space model,then solving quadratic programming problems,get optimal classification function(optimal hyperplane),finally,generate vector of the text to be classified into the model for category function ,according to the value we calculated to determine the category of the text.If the training set could be correct linear segmentation by a hyperplane ,And the distance between the nearest vectors with hyperplane are maximum,then this hyperplane as the best hyperplane.

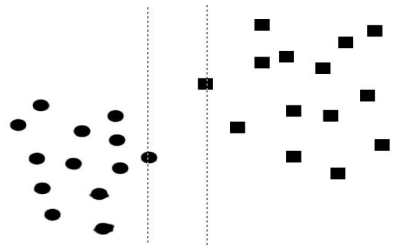

Figure 3. Non-Optimal Decision

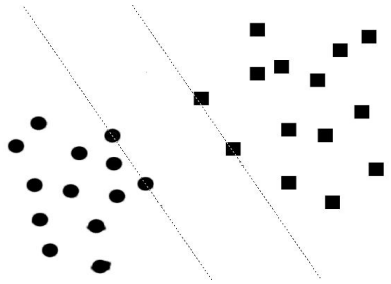

Figure 4. Optimal Decision

3.5.2 Topic Tracking Technology

Topic tracking is when a new web text appears,first,preprocessing the network text ,Compare it with the different categories that we calculated from discovered the topic .the main work includes.

a).After the emergence of the new text, preprocessing the text,Selection of feature items, weight calculation and so on ;

b).Compare the similarity between the new text and the previous classification model, Within a certain threshold, Division of text.If not set the threshold range,set the text to a new topic

c).Form a new text classification model.

For the text preprocessing, similarity search, feature selection, weight calculation, have been introduced in 3.3, 3.1, 3.4.

4 Conclusions

With the rise of social media, the number of netizens in Tibet grows significantly, more and more professional researchers pay attention to the public opinion generated from social applications.This article summarizes the process and the difficulties of network public opinion analysis techniques by consult the large number of documents,explain technologies on tracking Tibetan-Based network public opinion topics by features of Tibetan language and grammar,the situation of tracking Tibetan-based network public opinion in China and abroad,text pre-processing,text representation,text similarity calculation,text eigenvalues selection,calculate eigenvalues weights,classical text classification algorithm.Technologies on tracking Tibetan-Based network public opinion topics is still in the development stage,in the future also need to continually improve the classification model,more precisely on the track and public opinion research.

Acknowledgment

This work was supported by the National Key Technolog y 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]MENG Xiang-he, *Key Technology Research on* Tibetan Websites Topic Detection and Tracking[D]. Gansu:Northwest University For Nationalities.5-6,April (2013).

[2]WAN De-wen,Study on Tibetan Information Retrieval&Search Result Clustering and System Implementation[D]. Sichuan: Southwest Jiao Tong University, 7,(2013).

[3]LI Ai-lin ,The Research of Tibetan Text Classification Algorithms for the Analysis of Network Public Opinion[D].Gansu: Northwest University For Nationalities, 6,(2014).

[4]E.Bonabeau,M.Dorigo,and G. Theraulaz. Swarm Intelligence:From Natural to Artificial Systems[M].New York,Oxford University Press,(1999)

[5]D.J.L.,Goss5,FrallksN,etal. The Dyamies of CoLIeetiveSorting:Robot-Like ALLt and Ant-like Robot.Proceedings First Confereneeon Simulation of Adaptive Behavior;From Animals to Animates[C]. Cambridge,MA:MITPress,(1991)

[6] WU Yu,Research and Implementation of Key Technologies in Network Public Opinion Analysis[D].Sichuan: University of Electronic Science and technology,4,(2011)

[7] KANG Jian,Research on Multi-Agent and Swarm Intelligence Tibetan Network Public Opinion Management[D].Sichuan:Southwest Jiao Tong University, 3,(2015)

[8]JIA Yang-ji,LI Ya-chao,ZONG Cheng-qing,YU Hong-zhi A Hybrid Approach to Tibetan Person Name Identification by Maximum Entropy Model and Conditional Random Fields[J]. Journal of Chinese Information Processing,8,1,108-110,(2014)

[9] WAN Yuan Research on Mining of Internet Public Opinion Based on Semantic and Statistic Analysis[D].Hubei: Wuhan University of Technology, Ⅱ ,(2012)

[10] HAN Xiao-bin Design of Hot Event Detection System in Tibetan Web[D].Gansu: Northwest University For Nationalities, 10-11,(2012)

[11] JIANG Tao Study on Hot Topic Detection Based on the Analysis of Tibetan Public Opinion[D].Gansu: Northwest University For Nationalities, 13-15,(2010)

[12] LI Ai-lin The Research of Tibetan text classification algorithms for the Analysis of Network Public Opinion[D].Gansu: Northwest University For Nationalities,23, (2014)

[13] LUO Jie,CHEN Li,XIA De-lin,Research on Fast Text Classifier Based on New Key Words Extraction Method[J].computer application,04,32-34(2006)

[14] LI Ai-lin The Research of Tibetan text classification algorithms for the Analysis of Network Public Opinion[D].Gansu: Northwest University For Nationalities,24, (2014)

[15]KANG Jian Research on Multi-agent and Swarm Intelligence Tibetan Network Public Opinion Management[D].Sichuan:Southwest Jiao Tong University, 6,(2015)

[16] JIANG Tao Study on Hot Topic Detection Based on the Analysis of Tibetan Public Opinion [D].Gansu: Northwest University For Nationalities, 20-21,(2010)

[17]MENG Xiang-he Key Technology Research on Tibetan Websites Topic Detection and Tracking[D]. Gansu: Northwest University For Nationalities, 10-11,(2013)

[18]LI Xiao-di Research and Application on the Technology of Web Text Mining[D].Beijing: Beijing Jiaotong University, 16-17,(2015)

[19] LI Dan The Study of Chinese Text Categorization Based on Naive Bayes[D].Hebei: Hebei University, 17-18,(2011)

[20] LI Xiang-dong,XU Peng,HUANG Li,SHEN Xiang-xing,Research of Journals Manuscript Categorization Based on KNN Algorithm[J]. The Library and Information Knowledge ,71-75,04,(2010)

[21]Chen, Chunan, Weiwei Sun, Baihua Zheng, Dingding Mao, and Weimo Liu. "An incremental approach to closest pair queries in spatial networks using best-first search." In International Conference on Database and Expert Systems Applications, pp. 136-143. Springer Berlin Heidelberg, (2011).

[22] CUI Jian-ming,LIU Jian-ming,LIAO Zhou-yu Research of Text Categorization Based on Support Vector Machine[J].computer simulation,299-302,30(2013)