Research on the Extraction Technology of Hot-words in Tibetan WebPages

Chang-Zhi WANG, Gui-Xian XU*and Hui WANG

Information Engineering College, Minzu University of China, Beijing, 100081, China
** Corresponding author, email: xuguixian2000@bit.edu.cn*

Abstract: The construction of Tibetan corpus is the field of Tibetan information processing of basic work. This paper uses the technology of web crawler and pretreatment and real-time acquisition of web sites to obtain a large number of Tibetan corpus in short time. The hot words reflected the hotspot of Tibetan people's attention in a certain period of time. The paper draws lessons from the TFIDF for Tibetan text information extraction and the words of different locations are given different weights to extract the hot words. It is really effective to realize the construction of the raw Tibetan corpus and the extraction of the hot-words by self-made software.

1 Introduction

With China's reform and opening up, Tibetan regions has witnessed a rapid development. China has strongly advocated the construction of information technology so that the Internet penetration rate in Tibetan areas is increasing year by year, the number of Tibetan Internet users is also growing exponentially, as information sites using Tibetan as the main language become more and more, Tibetan information generated daily on the Internet are beyond count. As a vocabulary phenomenon of the Internet age, hot words reflect hot topics and livelihood issues of a country and people of the region in a period .Hot words have the characteristics of the times, and this reflects immediately. How to extract the Tibetan information effectively and hot words is very hot topic of worthy study.

At present, both Chinese and English information researches techniques have achieved good results, but the researches on Chinese minority languages are in the primary state. For the past few years, Tibetan and other minority language website have witnessed a rapid increase, which provides the study of minority language with sufficient materials. The Tibetan corpus is important data resource of Tibetan information processing [1], we can summarize, analyze, generalize, extracted relevant knowledge and information from large-scale Tibetan corpus. Rapid identification and directional tracking for hot words [2], we can quickly understand the people feelings, know the social dynamics and development trends, faster and more comprehensive grasp the trend of public opinion, thereby performing the correct guidance of public opinion and propaganda.

2 Background

In the corpus construction [3] and hot words extraction, the traditional way of corpus construction is 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 generally not large, manual work is too much, the cost is too high, the cycle of construction is too long, so that it cannot be timely updated corpus. [4].As Web2.0 technology becomes more and more mature, everyone is content creator, and 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, as the foundation of natural language processing research. Usually use the web crawler [5] to crawl on the Internet to grab data. Web pages are very blended got through the crawler, extracting effective information from the web page, mainly based on visual features [6], DOM tree [7], text features [8] and other methods of text extraction. As the acquired raw corpus, we use structured XML to preserve. In the specific operation, the majority of researchers use DOM4J and JSOUP to preprocess the web page.

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, filter stop words, count frequency of words and do other processing steps. Researchers assessing hot words are mostly based on the frequency of the hot words and historical frequency

fluctuation. Some scholars put forward different weights according to the position of word, which is one of the schemes for extracting hot words.

This paper will introduce how to use web crawler to excavate Tibetan related sites, structured process the acquired resources, and store as raw Tibetan corpus. And then carry on Tibetan text pretreatment on the structure of the raw corpus, and go on research of hot words extraction, hot topic tracking based on the corpus of features including the time, the source and the author and others

3 The proposed method

3.1 Information Gathering

Information collection is the first part of the whole project. Hot words extraction needs enough material, while the way of manual acquisition cannot meet the needs of research obviously. Therefore, it needs to get a lot of Tibetan corpus by web crawler. Here we use the Crawler4j open source crawler to obtain the data. Crawler adopts the Breadth-First strategy, and the idea is that the initial URL is highly relevant with the theme of web page in a certain range, and is highly fresh.

3.2 Preprocessing

The set of acquired original web pages may contain a large amount of information that is not related to the content of the text, such as the HTML markup language of web pages. These interference information is called noise. Removing web noise is very important for the work of the system. After denoising, the system can improve the reliability of the results, and simplify the structure's complexity of web tab and reduce the page size significantly, thereby reduce the spending on the time and space in the subsequent processing. In recent years, the technology of web page pretreatment becomes more mature. The web pretreatment technology mainly includes: the page deduplication and denoising

About deduplication technology of the web page, Border proposed shingling algorithm and Charikar proposed random mapping method based on the word [9].These are the two current mainstream algorithms:the complexity of the method shingle time is lower, while the accuracy of the algorithm based on random mapping is higher. About page denoising technology, there are three methods, one is based on the structure of the page, other one is based on template and the last one based on visual information.

To improve the pretreatment's efficiency and resource's utilization rate, it requires special analysis of each website page structure and then set a specific extraction rules program due to different Tibetan website page structure varies.

3.3 Word segmentation and Remove stop word

Mainly based on dictionary, semantic and statistic, Chinese lexical analysis has matured and every technique has its merits and demerits[10]. With the deeper research of Tibetan information processing, after decades of research, Tibetan text automatic words segmentation technology also made good achievements, some scholars have realized an automatic Tibetan segmentation scheme based on case-auxiliary words and continuous features [11].

About the Tibetan removing stop words [12], in order to enhance the effectiveness of the extraction of the hot words, for modal particle, adverbs, prepositions, conjunctions, itself has no clear meaning, only to put it in a complete sentence have a certain effect, such as the common "of" and "in". We screened the high frequency of the Tibetan vocabulary then these Tibetan words were sorted into stop word list.

3.4 Hot words Extraction

After word segmentation and removing stop words, count multi-frequency data, which means that it not only needs to count frequency of a word using in different locations in an article, but also needs to count the total frequency of collected corpus appearing in a certain time period?

Hot words extraction algorithm draws feature extraction of TFIDF [13], and then give word strings different weights according to different locations in the article,and give double weight to the word strings that appears in the title .

Obtained data adopts UTF-8 Unicode. After segmenting the data of one day and removing stop words, it forms a large table named P, in which each word C points to corresponding weight value: weight(c).

The statistical algorithm of string frequency and weight about strings of the length N is as follows:

Input: L preprocessed articles.

Output: P table

1. Extract strings in L articles. Filter stop words, generate table P containing N strings, at this moment, the weight is initialized to 0.
2. Generate title table ,which total frequency is T1 ,and content table, which total frequency is T2(total frequency contains repetitive word count),in which each word C is corresponding with frequency value, such as title _tf (C), content _tf (C)
3. Generate article table recording the number of articles that the word C appeared, in which each word corresponds with a value of DF (Document Frequency) such as article_ df (C).
4. for i=1 to N
5. if title()!=null then
6. output title_ tf ( )
7. else
8. output 0
9. end
10. if content()!=null then
11. output content_ tf ( )
12. else
13. output 0
14. end
15. output article_ df ( )
16. weight()

17. end
18. output P table

The final weight calculation formula is as follows:

weight () =log ()*log ()(1)

In fact, it adopts natural logarithmic function. After calculate the heat weight of above the hot words, sort hot words on the foundation of the weights in table and then extract hot word according to the set number.

4 Experiment

We set the range of hot words for a day or a week, data originates from mainstream websites of Tibetan information. As it is rare to find Tibetan forum information websites, there isn't information acquisition involving this aspect.

We collect all valid text page from the six sites in recent years. Here sets crawling range in several popular Tibetan websites. For example: 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. Then preprocess and organize into 38,768 XML file raw corpus, and then extract hot words from the XML files, after pretreatment, save an article as shown in figure 1:

Figure 1. Format of Corpus Text Saving After Preprocess

Then we selected the Tibetan information website news published on December 4, 2015 toextract hot words, totally grab 96 valid text files in this day. Thedistribution of the 96 pages is shown in figure 2:

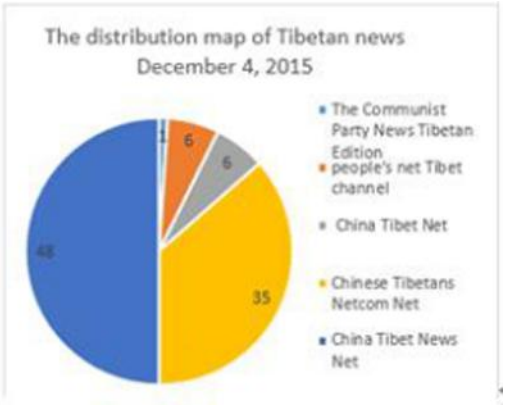

Figure 2: Distribution Map Of Tibetan News Text

The main program is shown in figure 3:

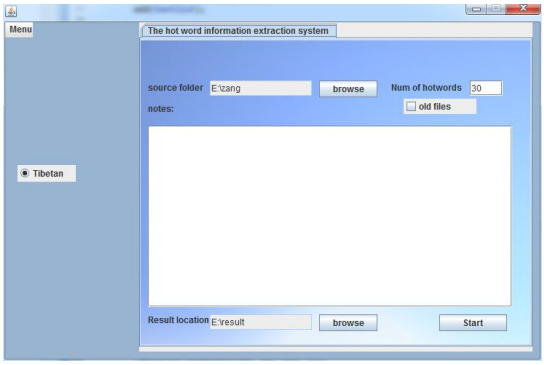

Figure 3: The Main Interfaceof Program

The results of the processing are shown in figure 4.

Experiments show that: applying the web crawler can effectively obtain all the corpus of news from Tibetan website, capture in real time, and analysis. Algorithm mentioned in this paper is effective to obtain the Tibetan hot words. As obtaining Tibetan websites almostbelong to the official website, the hot words distribute in multiple fields of politics, economy, culture, health care and other, and it reflects the main direction of publicity of Tibetan official website.

Hot-wo	TF	weight	DF	English
	121	0.01268027	23	China
	114	0.01024214	24	nation
	131	0.01007369	31	country
	68	0.00969097	7	Africa
	68	0.00951198	8	Poverty alleviation
	92	0.0092944	17	cooperation
	141	0.00925053	40	The masses
	57	0.00917713	6	The Panchen Lama
	126	0.00855953	42	city
	87	0.00833878	21	Culture
	71	0.00828941	14	medical care
	66	0.00826284	12	unite
	44	0.00769701	5	Tibet
	38	0.00694586	3	operation
	44	0.0064345	8	Changdu
	34	0.006398	3	event
	64	0.00636982	21	activity
	93	0.00604528	41	Sociology
	59	0.00593786	17	propaganda
	38	0.00590301	7	Xi Jinping
	27	0.00590148	2	petroleum
	38	0.00576261	7	Religion
	26	0.00569813	2	transfer
	61	0.00532392	23	Economics
	35	0.00518429	6	hospital
	35	0.00506051	7	Tourism
	36	0.00500864	9	education
	40	0.00499437	11	buddhism
	70	0.00481832	36	the people
	19	0.00478526	1	Yak

Figure. 4: Results of Hot Words Extraction

5 Conclusion

In this paper, software constantly obtains relevant Tibetan corpus by crawl on the multiple mainstream Tibetan websites. It grasps the main news material stored as structured Tibetan corpus through the relevant web information acquisition technology. After processing corpus segmentation and elimination of stop words, the final weight can be calculated by means of giving different weight to different words according to their places, as well as the word DF value, and hot words ranking can be drawn by ordering. The hot words are extracted by this method, which is simple and fast. Because of the lack of unified evaluation criteria for the extraction of network hot words, the accuracy of hot word recognition cannot be evaluated. The experimental results show that the method has higher accuracy. This paper provides a corpus of Tibetan information processing technology in this way. The classification of the hot words does not contain the parts of speech and recognition of some Tibetan names, Tibetan place names. There is no consideration of the historical frequency fluctuation of words, so the follow-up study needs to take it into consideration.

Acknowledgment

This work was supported by the National 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. DG Gao, B Guan. Retrospect on the Development of Tibetan Information Processing Technology [J]. Journal of Tibet University (2009)24(3):18-27.
2. YQ Li, LH Sun. Hot-Word Detection for Internet Public Sentiment [J]. Journal of Chinese Information Processing (2011)25 (1): 48-53+59.
3. DGGao, Tashigyal, DCZhao. Data Analyses of Large Basic Tibetan Corpus [J].Journal of Northwest University for Nationalities (Natural Science) (2013)34(92):46-51.
4. PFLi, QMZhu, PDQian. Construction Approach of Large-scale Corpus Based on Web [J]. Computer Engineering,(2008)34(7):41-46.
5. DZYang, GZhao, TWang. Application of WebCrawler in information search and data mining [J]. Computer Engineering and Design, (2009)30(24):5658-5662.
6. QWu, XYang, ZXZhao. Web information extraction based on visual characteristics [C], Symposium of the Sixth China Conference on Information Retrieval (2010).
7. RXZhang, MQSong, YLGong. Parsing DOM Tree Reversely and Extracting Web Main Page Information [J]. Computer Science. (2011)38(4):213-215.
8. JD Hu. Research on Web News Extraction and Duplicates Elimination [D].ZHEJIANG UNIVERSITY (2011).
9. CQMa, XGMao. Research on Near-duplicate Detection Algorithm Shingling and Simhash [J]. Computer & Digital Engineering(2009)39(1):15-17
10. JW Mo,Y Zheng, ZY Shou, SL Zhang.Improved Chinese word segmentation method based on dictionary[J].Computer Engineering and Design.(2013)34(5):1802-1807
11. YZChen, BLLi, SWYu, CJLan.An Automatic Tibetan Segmentation Scheme Based on Case-Auxiliary Words and Continuous Features [J]. Applied Linguistics(2003) (1):75-82
12. JZhu, TRLi.Research on Tibetan Stop Words Selection and Automatic Processing Method [J] Journal of Chinese Information Processing(2015)29 (2):125-132
13. CYShi, CJXu, XJYang. Study of TFIDF algorithm [J]. Journal of Computer Applications (2009)26:167-170.