Analysis and Countermeasure Research about Sensitive Information of Complex Cyber Psychology of Cross-border Multi-ethnic Language

Yingyu Yang[1, 3, a], Bingze Li[3*], Jiamei Wang[2, 3], Changsen Yuan[2, 3], Feng Liu[2, 3], Gang Hu[2, 3], Guangming Lu[2, 3], Rui Lin[2, 3]

[1]The University of Bath, the United Kingdom;

[2]School of Electrical Information Engineering, Yunnan MinZu University 650500, China;

[3]Yunnan Province Colleges minority language information processing engineering research center, 650500, China.

[a]1575763065@qq.com

Abstract. In this paper, we consider all aspects of sensitive information of cyber psychology of cross-border multi-ethnic language, and use the method of combining the theoretical research and empirical research to carry out the analysis and countermeasure research about sensitive information of complex cyber psychology of cross-border multi-ethnic language in Yunnan province. It closely combined with the practical application and social needs and it also provide technical support and method guidance for sensitive information mining of complex cyber psychology of cross-border ethnic language. Due to China as a multi-ethnic multilingual country, so the development value, the demonstration value and the theory value of this topic is worth attention.

Keywords: Cyber Psychology; network language violence; Content analysis.

1. Introduction

In recent years, the sensitive information of cyber psychology of cross-border ethnic areas threat the social public security in the form of "content threat". At present, study of security respect of sensitive information of cyber psychology of ethnic language is still in its infancy. Though some scholars have begun to step in this field and published a number of influential achievements. But we also lack of discipline system support and the results of comprehensive system. Therefore, it is important to innovative study the content of sensitive information of cyber psychology of cross-border ethnic language in Yunnan province. How to combine with related factors about aboriginality and differences which is regard to economic and social development in ethnic regions. It is also important. It is of great significance to promote cross-border economic and social development in ethnic areas, purify network environment, maintaining border security and stability in ethnic minority areas, make up for the flaw of network security monitoring system in our country, further the study of language and semantics, enrich public opinion monitoring theory. In view of the diversity and complexity of cross-border ethnic language in Yunnan province, this topic through a case study as the breakthrough point, so we select one of the cross-border ethnic groups in Yunnan province as the research object--the Yi nationality. We research and design the key technology of sensitive information of cyber psychology of cross-border ethnic language in Yunnan province from the domestic and foreign new theory, new method and new technology in both English and Chinese research present situation.

2. The key technology research about sensitive information mining of complex cyber psychology of cross-border multi-ethnic language in Yunnan province

The research method of this paper is combining theoretical simulation and experimental measurement, and this paper through the analysis of the experimental results further design optimization theory, laying a solid foundation for the system of practical application. The implementation of the project is divided into two parts: The first part is the technology research about collecting and grabbing the network public opinion information resources of the Yi nationality; the

second part is the technology research about results of the analysis. As shown in figure 1, it is the key technology about sensitive public opinion information mining of cyber of multi-ethnic language.

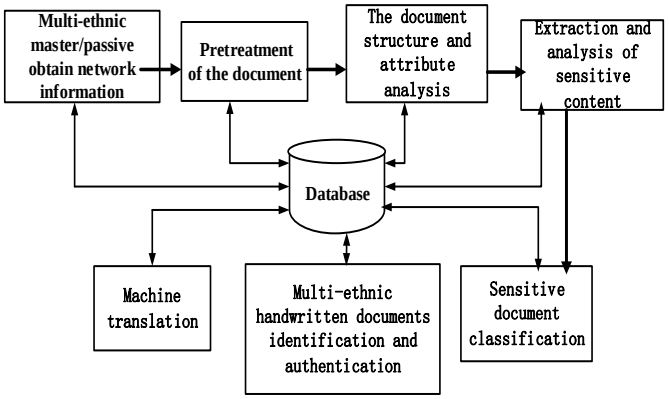

Fig. 1 the key technology about sensitive public opinion information mining of cyber of multi-ethnic language

2.1 Technology research of collecting and grabbing the network sensitive information of the Yi nationality.

Network sensitive information acquisition, content filtering, has become a hot research topic in recent years, it is not only has great practical significance and it also important way for accessing sensitive information of cyber of cross-border ethnic language. It is worth further study. In addition to network bottom analysis and research, we usually use web crawler to implement. Chinese web crawler technology is a mature technology. Now common search strategies are: Based on the IP address, breadth and depth of preference, the best first search strategy. Many researchers put the breadth-first search strategy in topic crawler, they think the strategy and the initial uniform resource locator on certain links within web has topic correlation probability is very big. In order to meet the needs of application of personalized and improve the efficiency of the crawler, topic crawler arises at the historic moment, the target is extracted from a lot of web pages associated with a particular topic or page, will filter out irrelevant links and page. But the web crawler technology of national characters is still in the phase of basic research. In the literature [1], they got a brief discussion of designing and accomplishing Uygur, Kazak, and Ke multi-language search engine, accounted ethnic minority languages existing webs, and design initial URL addresses.

In this paper, the basic task of the information collected is collecting rich, all kinds of public opinion information from a variety of data format page. It provide the required data for public opinion information mining layer, is the precondition of sensitive public opinion deep mining. The basic process of collection is as follows:

In information acquisition target website, the basic process of collection is as follows: (1) The crawler management module mainly use the web crawler technology to obtain Internet public opinion information, its basic principle is: Order to read a list of URL, access the page which is URL points to through protocol on Web, and then extract the new URL into the URL in the queue from page has access, and search information according to certain strategy. Due to the different Internet web page structure, various scripting languages, data format and a single web crawler is difficult to meet the network public opinion analysis system for the needs of different web page source data acquisition. Therefore, this article use the way of topic web crawler and collect their seed URL address, nature of the web site, the used information of the character encoding standard, obtain network information from the given URL list, that is counted web site of Yi language word building. The beforehand URL is established set as a crawl queue. On the basis of general crawler strategy (breadth-first crawler algorithm), considering the particularity of minority languages web page, combine the analysis of the text, design the correlation measurement method of the sensitive topic, collection for minority theme page, and, to extract only links to the page text in the link analysis and extraction, and to other resources (images, audio, animation, etc.), all links to be filtered out, when it is necessary we should consider the code cleanup and code conversion.

Research the sensitive network public opinion classification technology. There are two kinds of text classification methods, The first method is based on knowledge, such as Reuters classification system and Chinese library classification, The second is based on statistical method, decision tree and naive Bayesian, artificial neural network and SVM, and The SVM is currently recognized the most effective machine learning algorithms for text classification. The method based on knowledge is suitable for a particular field, and to define the inference rules for each category according to the knowledge of this field, we can see a document whether meet the above rules to judge whether this document belong to this category. Based on statistical method does not need complex domain knowledge, but rather through a large number of observed similar document and sum up experience, as the basis for the classification of same document. This method has good efficiency and overcome the narrow method of based on knowledge is now widely used.

2.2 Researching the Result of Analysis

Yi language sensitive phrase in text message flow often can reflect the hot topics and emergency in the message flow. In this paper based on the law of the sensitive Yi language phrases and the analysis of the existing technology, in view of the huge challenges about diverse of sensitivity measurement method number and huge number of network text messages, network sensitivity measurement, description requirement based on the calculating frequency sensitivity of phrases in history, give the specific definition of high-speed Yi language text message flow sensitive phrase mining tasks. After that we adopting a high accuracy but expensive simple algorithm in space and time, can record the frequency of each phrase in history. And update the method by the method of sampling the dynamic message flow. based on the analysis of several features about text message flow in the network, we can determine phrase sampling time through the estimate the frequency of missing phrases appear dynamically.

3. Design and Implementation of Yi Language Web Page Sensitive Information Mining Filtration System

3.1 The determination, remove duplicate, extraction, storage of Yi language web pages

This paper use c + + programming language, Microsoft Visual Studio 2010 software development platform and ACCESS database to development and design. It mainly divides into four parts: determine parts of Yi language web sensitive information filtering system, web page remove duplicate parts of sensitive information filtering system, web page extraction parts of sensitive information filtering system, and storage parts of sensitive information filtering system. Basic idea is: design of Yi language network sensitive information filtering system to obtain related algorithm and crawler, using the determination algorithm to collect related Yi language web page. Extracting some of the key information of S web page sensitivity information filtering system, the main content of the extracted pages include the text of the title, text, date, and link (URL), and other key information, and deposited in the database; To facilitate the next step of Yi language data analysis work, we extracted text of the Yi language sensitive information filtering system and after judging encoded converts into a unified code then stored it, we chose Chinese Yi web as an example in this paper, research access technology of network sensitive information filtering system. This article selects the Chinese Yi web Yi language version as a web page collection objects, scraping and collecting Yi language web, the homepage URL address is: http://222.210.17.136:81/zgyx/indexyi.html.

Determination of sensitive pages: First, acquire the homepage address content, second, acquire the entire effective URL on the page and add the URL to the database. Part of the code is as follows:

```
string html = http.GetHtml(url);
List<string>urlList = getUrlList(html);
int Count = AddUrl(urlList);
```

Before collecting web content of sensitive information, need to filter the saved URL if the article page URL links, and decide whether the URL has been collected. Part of the code is as follows:

```
if (OldUrl.url.Contains("/details_yi.jsp?") == true)
{
```

```
List<Article>aList = DbFactory.getArticle("select * from [article] where url='" + OldUrl.url +
"'");
    }
```

Sensitive web page extraction parts: judge and extracted sensitive to the title of the page, publishing time, source, URL, text information by parsing the Yi language page information acquired and transform sensitive information extracted format then stored it in the unified ACCESS database, to facilitate the next step to retrieve related work. Figure 2 are the store information of ACCESS database.

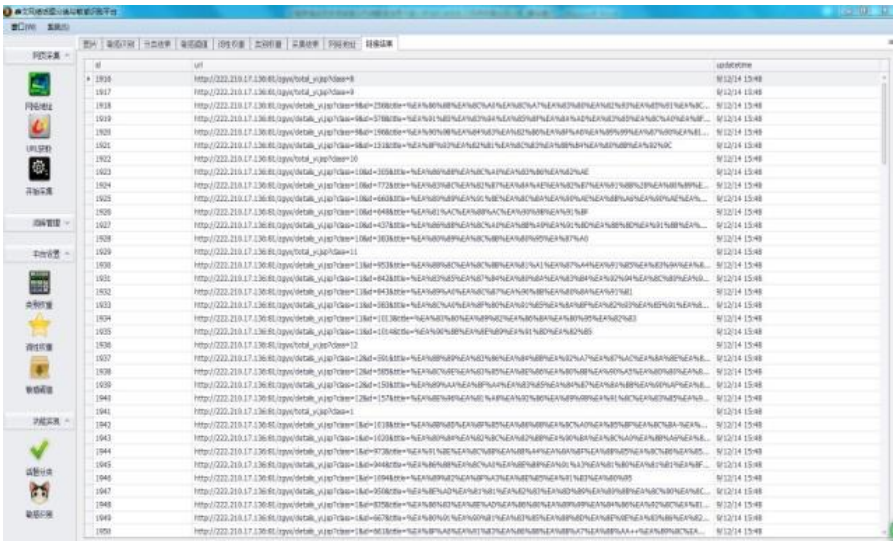

Fig. 2 the store information of ACCESS database

There are many indicators about testing and evaluation in web crawler acquisition algorithm. There are two important indicators: recall ratio and precision ratio.

(1) Recall ratio: R= the amount of related documents collected/ the amount of related documents in system*100%.

(2) Precision ratio: P= the amount of related documents collected/ the amount of related documents in system*100%.

In this article, the range of the crawler fetching limit in Chinese Yi web range, so in the performance test, the amount of related documents of web equivalent the total amount to document. Therefore, two indicators are same; here we use the recall ratio.

Through the collection platform of Yi language web page information, the totals of URL collected are 92 in China Yi web, that is the amount of related documents in system are 92; text document number Stored in the TXT is 62, which is the amount of related documents collected is 62. Therefore:

Recall ratio= precision ratio=62/92*100% = 67.39%

The test result shows that when we collect Chinese Yi web (Yi language version), collection and accuracy is not high, through the analysis we found that the reason is mainly due to image format and the Yi language website of the part of the Yi language web page is too little, this led to the collection and the result is not ideal.

3.2 Analysis and Recognition Results of Yi Language Web Page Sensitive Information

In this paper, we encounter difficulties of cross-border ethnic network public opinion information mining in Yunnan in the each link; we chose Yi language as an example to further research technology of Yi network public opinion information mining. We obtain the good effect by three aspects: further research, research and system design and implementation of collection technology about Yi language network public opinion sensitive information, network public opinion classification technology and sensitive topic identify. Figure 3 is renderings of public opinion weight category set, figure 4 is the sensitive information recognition results, and figure 5 is sensitive information classification results, specific as follows:

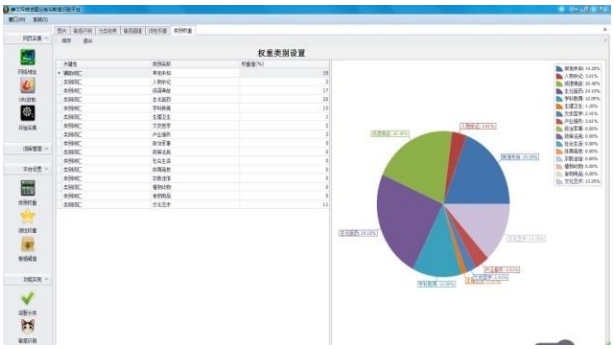

Fig. 3 Renderings of public opinion weight category set

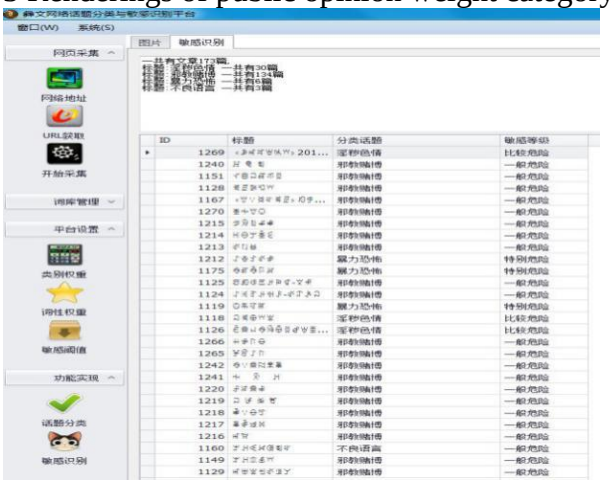

Fig. 4 Sensitive information recognition results

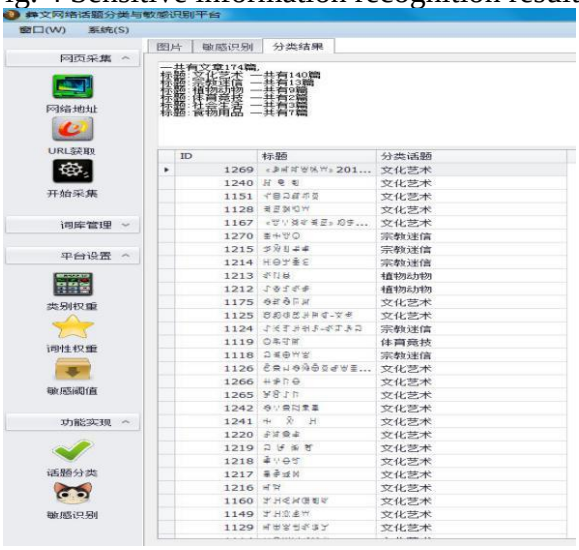

Fig. 5 sensitive information classification results

4. Conclusion

Now personal psychological information becomes the target of a lot of criminals, there is a security risks that is not optimistic, if the lack of protection strategy of information security and technical support, it will cause personal sensitive information leakage. In this paper, we analyzes the problems of personal sensitive information leakage of the Internet according to the situation of Internet personal psychology sensitive information protection security, discuss the protection measures of personal psychological sensitive information on the Internet from the aspects of internal and external security on the Internet, and maintenance the Internet personal sensitive information security.

In this paper, we chose Yi language as an example, analysis of sensitive information of complex cyber psychology of cross-border multi-ethnic language in Yunnan province. In this paper, we

reference the Chinese method and the Yi language itself is different with Chinese, so there is certain limitation in the Yi language web page sensitive information identify areas, such as recall ratio and precision ratio is not very ideal, it also need to further research. Especially there are certain limitation in the semantic expression of Yi language text characteristic and the use of standard sensitive words library etc. such as the aspect of Yi language standard sensitive words library development ways, there are more development and standard word library in both English and Chinese we can use, for the Yi language, there is not a standard set of sensitive word library. In this paper, we filter out the sensitive information text and complete the risk classification through the development of Yi language standard sensitive word library, part-of-speech tagging, Chinese and Yi language translation thesaurus, Chinese and Yi language sensitive level in the thesaurus building, it can only analyze performance objectively, and it is unable to compare, but the development value, demonstration value and theoretical value of this topic is worth to attention for the first people who attempt to Yi language web sensitive information identification, for China as a multi-ethnic multilingual country.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (61363085), Major projects in Yunnan Province (ZD2013013), High level construction of university scientific research project of Yunnan MinZu University. Yuan said information and extraction method of the minority language and culture"(WT125-61).

References

[1] WU Lihui, WANG Bin, YU Zhihua, Design and Realization of a General Web Crawler [J]. Computer Engineering, 2009, 31(3): 123-124.

[2] Sili Wang, Research on the Automatic Discovery and Collection Technology of Tibetan Web Pages [J], Northwest University for Nationalities2010.

[3] Carlos Cobos, Henry Munoz-Collazos, RicharUrbano-Munoz. Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion. Information Sciences 281 (2014) 248–264.

[4] Wang JiaMei. Yi Ancient Digital Research - Yi Ancient special website. College. 2006.07. 149-150.

[5] Li XunShan, Wang JiaMei (Corresponding author).Research on Yunnan Standard Yi Language Fonts Design and Character Coding ,Electronic Science and Technology, 2011.24 (5). p97

[6] FENG Hao, WANGHui, WANGJia-mei (Corresponding author).Design and implementation of Yi character input method based on free split mode. Journal of Computer Applications.2010.A01, 306-308.

[7] FENG Hao, ZHOUYing, WANG Jia-mei. Popularization and Application of the Yi Character Input Method's Free Split Mode. Modern Computer, 2011.9.3-5, p11.

[8] CHEN Shun-qiang, ZHANG Yang, XIONG Jian. Sichuan ancient Yi language fonts design and character repertoire code .Journal of Southwest University for Nationalities (Natural Science Edition), 2009, 35 (4): 913-918.

[9] PU Zhang-kai. A Dictionary of South Yunnan Yi. The Nationalities Publishing House of Yunnan.2005.10.

[10] Yi character set of Yunnan, Sichuan, Guizhou and Guangxi. Published jointly by three Nationalities Publishing House of Yunnan, Guizhou and Sichuan.2001.

[11] ShaMaLayi. Development of computer operating system of the Yi language [J].Journal of Southwest University for Nationalities (Natural Science Edition), 2003.29 (1): 1-8.

[12] ZHU Jian-jun .Some Thoughts on the Setting up an Ancient Yi Character bank. Journal of Huzhou Teachers College [J].2003.25 (1):22-25.

[13] Li Jinfa. Exploitation of Yi Character Coding Conversion Software. Journal of Yunnan Nationalities University (Natural Sciences Edition), 2008.14 (1).

[14] Fu Xianghua. Design and implement the IOB operation platform and IOB database system. Master's thesis. Northwest A&F University.2002.

[15] Fang Ziwen. Design and implement the Ancient Yi Windows input platform. Master's thesis. Yunnan Nationalities University.2008.5.

[16] Lu Jixing, JiangWeike, Zang Yueli, QuBin. Input method based on Windows IME [J] Journal of Agricultural University of Hebei, 2003.26 (5): 290-292.