Design and Implementation of Network Public Opinion Monitoring and Analysis System

Dezhi An1,2, Guangli Wu1,2, Jun Lu1,2, Shengcai Zhang1,2, Yan Li1,2
{ adz6199@gsli.edu.cn}

1 School of Information Engineering, Gansu Institute of Political Science and LawLanzhou 730070, China
2 Key Laboratory of Evidence Science of Gansu Province Gansu Institute of Political Science and Law, Lanzhou 730070, China

Abstract. Firstly in this paper to the domestic and foreign public opinion monitoring system and its application to the process of development and construction experience, the research results summarized based on, combined with current solutions and solutions, organize and optimize the main demand of public opinion system, system requirements are obtained, including data acquisition and pre processing needs, public opinion monitoring and analyzing the demand and application show the demand and management, business management requirements, system security and performance requirements. Secondly, the web crawler, text extraction, Chinese word segmentation and semantic analysis and text clustering techniques based on, from the characteristics of public opinion data, data acquisition, data transmission and storage mechanism, function realization characteristics are the public opinion monitoring and analysis system of the whole structure and the design idea, and from the infrastructure platform, data acquisition, data pre processing, a monitoring and analysis processing, web application of the five angles expounds the function module of the system design scheme, at the same time, the article also gives the system of business process analysis and optimization, deployment architecture and the way of realization.

Keywords: public opinion monitoring and analysis; data collection; data mining; text clustering; system design

1 Introduction

Public opinion is in a specific time and scope, based on the changes of some internal or external factors, and the formation of the concept of social people to change the idea of a collection. With the development of informationization in our country, traditional information dissemination mechanism and channels gradually let in Internet technology, especially in the Web2.0 after a period of time, all news website, stick, microblogging, forums and other among the people made widely recognized and applied, have also become the people to express the views of some event, views, aspirations and feelings many feedback of the main channel. This channel reflects the personal views, views, aspirations, attitudes, and many other collections, that is, the network public opinion. Due to the characteristics of the Internet itself determines the network public opinion can be through the related computer technology for data collection and analysis, in order to ensure the accuracy and reliability of the information

at the same time [1], improve the comprehensive information and the speed of response, which is the public opinion, the system provides a prerequisite. Under the current social conditions, the network has become an effective way to realize the citizens' right of supervision, and plays an important role in protecting the legitimate rights and interests of citizens, promoting the development of public welfare undertakings and maintaining social harmony and so on. In the face of public events, the network group has the characteristics of emotional, blindness and conformity, which makes it easy to use these features to spread the information and form a negative direction of public opinion. Thus, the network public opinion both advantages and disadvantages, only government to deepen the reform of the cultural management system, through the mainstream media reasonable guide public opinion, for networks with bursty sensitive events, and constantly improve the level of crisis public relations, in order to harm is reduced to a minimum. Therefore, the establishment of long-term mechanism of public opinion supervision, strengthen the communication of the government and cultural enterprises and institutions, timely detection of hot events on the network and sensitive topic, analysis and forecast to the network of public opinion, and the mainstream media for positive guidance, to the conversion of public opinion to promote social development and maintenance social harmony is one of the power.

2 Summary of Technology Theory

Search engine is the modern society fast access to one of the most effective way of information and global famous search engine Baidu and Google to search according to the key words, rapid search needed information, realize the sharing of resources. Web crawler is the main support means of Internet search engine. This program through in each information bit automatic crawl, from a specific URL of search page, according to the demand of the acquired information stored in the resource library, and a new page in the repeated retrieval, ad infinitum [2].

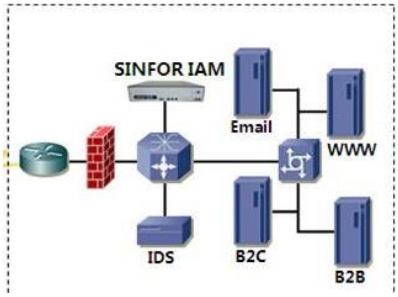

Fig. 1. Sketch map of web crawler

2.1 Web page Text Extraction Technology

Web interface is usually after the web crawler technology to deal with the resulting interface, generally Html text format. Web interface is carried by the open resources, mainly from two aspects: first, the portal site to express the theme, also known as web pages, such as video sites in the film and television resources, etc.. The two is not related to the content of the web page, such as the interface of the toolbar, properties and other noise information. Usually

web design is not uniform implementation of the standards, according to the requirements can be designed into a variety of styles, so there is no uniform layout of the web page text and web noise. In some website interface, network text and network noise layout is not reasonable, not only for primary and secondary user is difficult to distinguish to delete selected useful information, and will lead to some based on the analysis of the theme of the web crawler program work difficult and prolong the processing time, reducing the accuracy. Network information detection and analysis of public opinion, it is a requirement to effectively extract the content of the main body of the network, filter out the noise of the network technology. This technique uses a set of pre programmed algorithm, in a short period of time to lock the theme, and automatically shield off the non associated resources, is based on the Web network crawler work procedures in the important steps [3].

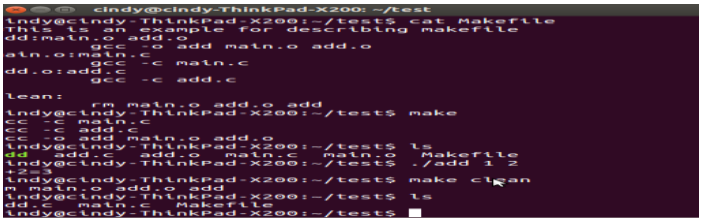

Fig. 2. Sketch map of web crawling experiment

2.2 Topic Monitoring and Tracking

Topic tracking technology is for massive Internet information database, according to the topic and the Internet has been the main information, automatic screening and detection of new topics and continue to track the technology. The technology using computer program, independent detection network, the current focus of the topic and issued warning information, and tracking and analysis of known topics, such as topic of time, development process and end results. Topic detection and tracking tracking technology are targeted to specific, usually focus events occurring within a certain period of time, such as hot news, sensitive events, contains the event itself and the associated with the event of other things, such as public opinion, social impact, derivative products. Topic monitoring and tracking technology usually consists of the following four steps: first, the new topic detection. Refers to the use of a specific topic detection program, in the Internet information flow intercepted a new topic, and the topic is classified into a known topic library, and to continue to track the topic. Second, the first time to report the detection task. Refers to the first report of a new topic, based on the sequence of events and the determination of the target. For the first time, the work reported by the first time is only the first time to detect the content, and the final test is not necessarily consistent, but to a large extent, convergence. Third, related testing tasks. Related testing refers to the program for any two web content of the content of the comparative analysis, through the correlation degree, to determine that they belong to the same topic of the possibility. Fourth, topic tracking task. Refers to the task of the topic has been clear to continue to track detection, to obtain the follow-up related content of this topic. The feature of this task is that there is no clear known topic, but by some previous related results to determine. Topic tracking after the target topic of all the content of the test, to determine the relevance of the topic and the topic, to complete the tracking task [4].

2.3 Text to Quantization Model and TF-IDF Algorithm

According to the way of content selection, the text is divided into three directions: Boolean, vector, and probability. Usually case, gist of text can be expressed in the keywords or index terms, index terms by the main content of the text summarized that, with the index, the reader can quickly overall understanding of the whole text. Vector model is using mathematical modeling methods, the various lexical set into a lexical space, information retrieval and text are the space vector, by calculating the approximation of the two kind of vector, lexical space approximate measure of the degree of text content of similar degree, for the classification. Vector method is divided into two kinds: generalized vector model is pointed out and keyword vector does not exist cross relation, but linearly independent, so between the keywords has interdependence, this dependence by the keyword synchronized to produce the model output.

3 Overall System Design

This chapter gives the overall structure and design ideas of public opinion monitoring and analysis system from the point of view of the characteristics of public opinion data, data collection, data transmission and storage mechanism, function realization and so on. Next, this chapter will explain the module function design of the system from five aspects of the infrastructure platform, data acquisition, information pretreatment, monitoring analysis and processing, WEB application. At the end of the chapter, the business process design and deployment architecture of the system is presented from the point of view of user's demand, data security and system performance.

3.1 Overall Design of the System

Public opinion analysis system based on the needs of the system is to collect various news websites, forums, micro-blog, Post Bar data from the network, simple filtration and re scheduling pretreatment, so as to realize the topic of surveillance and tracking, identification and screening of hot sensitive topics and other functions, and eventually fast WEB application show, which requires the system must be on the following issues: one is how to solve the problem of news website, forum, micro-blog, Post Bar etc. the different formats of data acquisition and pretreatment, the data acquisition and preprocessing, to unified format data and classified; the two is how to perform data mining and quantification of the text, so as to realize the the measure of a text; the three is how simple and efficient implementation of monitoring and tracking, hot topic selection, sensitive topic recognition and module function. By which design and algorithm is reasonable; the fourth is how to design web application to show the function, so that the whole system can be friendly and user interaction, how to design the data storage and transfer mechanism, ensure the rapid processing of data access behavior.

3.2 System Module Function Design

According to the overall design and system architecture of public opinion, based on the analysis of user requirement, the system is mainly divided into four functional modules: data acquisition module includes data acquisition configuration, URL filtering, URL re scheduling, acquisition control and data acquisition; data preprocessing module includes the functions of

noise elimination, text extraction, Chinese the segmentation and semantic analysis; public opinion analysis module including text to vector quantization and TF-IDF calculation, monitoring and tracking, topic selection, public opinion hot sensitive topic recognition function; WEB Application module including news public opinion, special topics, hot topics and sensitive topics such as the display function, and system configuration, information retrieval and security management functions [5].

3.3 System Business Process Design

Users of the system using public opinion monitoring and analysis system mainly for public opinion data query, thematic display and analysis, sensitive topic monitoring, public opinion focus and other business activities, and on this basis, to implement the public opinion information collection, analysis, storage, statistics application function. The user's basic business process is as follows: firstly, the management of the system or special maintenance personnel to log on to the system, the initialization operation on the data collection configuration information, sensitive keyword thesaurus, the relevant decision threshold, the target site and micro-blog Post Bar; secondly, with the corresponding user entered data acquisition function, select acquisition configuration set the URL filtering rules for each target site and re scheduling rules, and determine the frequency corresponding to each target address; thirdly, the system administrator to set the automatic execution of data acquisition, data preprocessing, user login to the WEB application interface with public opinion public opinion information permission to view monitoring and tracking of public opinion and view a specific topic, the public opinion hot or sensitive topics, and derives the related public opinion graphics or report page, people Work process.

4 Conclusion

This paper on the domestic and foreign public opinion monitoring analysis system development process, construction experience and research results are detailed and summarized, and combined with the current domestic public opinion monitoring solutions and solutions, many research and consulting, and strive to full consolidation and optimization of system requirements. The content of the paper mainly includes data acquisition and preprocessing, public opinion detection and analysis demand, application display and management requirements, business management requirements, system security and performance requirements, etc.

References

[1] Cao Shujin, Chen Yijin. Research on function and model of network public opinion information analysis and utilization. Journal of information resources management. No. 03. (2011)

[2] Anni Zhang, Jianghua, Hao Xianglian. Design and research. Huaiyin Institute of technology subject oriented fast search engine journal. No. 03. (2014)

[3] Gao Hui, Wang Vujacic, Fu Yan. A long-term trend forecast method of. Web public opinion Journal of University of Electronic Science and Technology. No. 03. (2011)

[4] Zhang Huanming. Research and design of network public opinion analysis system. micro computer information. No. 18. (2010)
[5] He Jia, Zhou Changsheng, Shi Xianfeng. Realization of network public opinion monitoring system. Journal of Zhengzhou University (Science Edition). No. 01. (2010)