TELKOMNIKA, Vol. 13, No. 4, December 2015, pp. 1368~1375
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i4.2811

Design and Implementation of Network Public Opinion Analysis System

Ma Junhong*[1], Liao Na[2]
Department of Computer engineering, Institute of technology, Xi'an International University
Xi'an, Shaanxi, China, 710077,
*Corresponding author, email: maxiaofei913@163.com[1], 2340869@qq.com[2]
Phone: 029-87866350[1], 029-88751119[2]

Abstract

Network public opinion analysis is an important way of information analysis processing. This paper based on the research of the related technologies, designs and realizes a new network public opinion analysis system. System mainly includes network data fetching part, fetching the data processing part, analyzes the processed data part and display part of the public opinion analysis results. In the document extraction part, used the web crawler technology, Larbin web crawler to realize the collection of web content; In public opinion information analysis part, the implementation of the new topic adopts an improved Single - Pass clustering algorithm. This algorithm is using of multi-center, using the title and body of the vector to compared two-way, that is better reflect the dynamics of public opinion topics. Finally, in the network environment of a university, we have the tests repeatedly. The results show that the new public opinion analysis system running is stable and has good efficiency. The thesis has certain value for the development of other information analysis systems in the Internet.

Keywords*: Network Public Opinion, Search Engine, Cluster analysis, Web crawler*

Copyright © 2015 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

Public opinion analysis is a method by collecting information, text crawls and other related technologies, to quickly find and gather relevant information on public opinion. Meanwhile, the information collected automatically capture, text filtering, topic analysis, text classification, clustering analysis and statistical judgment [1]. February 2015, China Internet Network Information Center CNNIC issued the "35th Statistical Report on Internet Development in China". T Report" shows that, with the rapid rise of the mobile Internet, the 2014 Chinese netizens had reached 649 million. Of which 54.5 percent of Internet users said trust information on the Internet, there is 43.8% of Internet users said they liked on the internet to comment. On the other hand, in recent years, the Chinese government actively promote and guide the network in politics, the majority of Internet users through the Internet channel comment on current affairs, reflecting the people's livelihood, suggestions, networking has become one of the most important platform for users to express their views, the emergence of a vast network of public opinion macroscopic power. At the same time, a growing number of Internet public opinion began to stand out events, and to bring a positive or negative social impact [2].

Colleges and universities are an important part of society, the network has become the first university students' access to information and media for communication, the public opinion, public opinion and social status situation has similarities [3]. College Network public opinion in favor of improving the management of universities working to promote democratic and harmonious, but also to all kinds of false statements, exaggerated speech, malicious speech provides a breeding ground, to scientific research, university teaching campus stability and harmony adversely affected. And because students blindly follow, concentration, as well as university network monitoring mismanagement and other reasons, making it easier to become a college cyberspace network public opinion of the outbreak, the social influence of the student population growing network of public opinion, for college and university students and stable growth influence also taught growing. Therefore, the study of the university network public opinion analysis system is particularly important.

Received September 7, 2015; Revised November 4, 2015; Accepted November 20, 2015

Foreign public opinion research is similar to the western public opinion research. From the beginning of the 19th century, the western public opinion study by social scholars, political and social psychologists and other social science and the wide attention of scholars, and rapid development. Related research mainly focuses on the government decision-making, borrow online poll, and as a reference for the policy [4]. Parties of all countries, it is through the network media the electronic Bridges, implementation of party members, the masses party directly or indirectly participate in making important decisions, which is conducive to better integrate the interests of all parties, improve the democratization and scientific process of decision making. In the university network public opinion management, study abroad is more focus on the school crisis management. With the deepening of the crisis management research and development, the school crisis management has become a new research field. Some European and American countries and Japan for the school crisis management research relatively early, has the certain data accumulation. American research project TDT (Topic Detection and Tracking) mainly related to the five areas of research: continuous text segmentation (for broadcast news), the theme track, theme discovery, discovery of new events, relevant findings [5]. Its intention is to come up with some algorithms that can discover and summarized from the data stream important information and content. With the deepening of the crisis management research and development, the school crisis management research has become a new area of research. Some European countries and Japan and crisis management school relatively early, there is a certain accumulation of information.

In recent years, along with our country for the work of network public opinion management, some only the network public opinion management laws and regulations, the research on network public opinion officials and private institutions support increase gradually, some rely on the government, the media public opinion monitoring agencies, academic institutions and research institutions arises at the historic moment [6], the network of public opinion, more and more researchers writings also more and more Founded in December 2005 the communication university of China institute of public relations and public opinion, is an analysis of the public opinion information research and academic research institutions, the main research interests include social public opinion, crisis warning, brand reputation, public relations activities, etc.; Renmin university of China and founder group jointly established "the National People's Congress a founder public opinion monitoring research base", there are some other colleges and universities set up the relevant research center. The establishment of the network public opinion agencies got the attention related to network public opinion. Xu Xin was proposed based on signal analysis of early warning mechanism of network, and put it into two kinds of modes: signals longitudinal excavation and lateral control. 2011, Hong XiaoJuan etc. in the “university network public opinion assessment system based on the model I-Space to build” had introduced the British economist bowie Sauter's "information space" (I-Space) model, Coding, abstract degrees and diffusion is the model of three dimensions. 2013, Pan Chao based in the croak of game of network public opinion and government regulation model "put forward in the connection between the network public opinion and supervision; Yao Chunhua in the study of network public opinion control technology based on weibo through the analysis of the characteristics of public opinion transmission in weibo, micro blogging public opinion control technology solutions are put forward [7].

Network public opinion analysis can help us to automatically collect the required data, found the problem and carry on deep mining, and then find the topic of the relationship between different factors, on the whole details of an event; this has important implications for network monitoring public opinion.

2. Design of Internet Public Opinion Analysis System

The basic principle of network public opinion analysis system implementation is to WEB pages of text information collection, processing and data mining [8]. Before data mining, there is the information preprocessing, namely WEB page filtering, text segmentation, word frequency statistics, feature selection and feature extraction, etc. System will eventually get hot issues arising from the recent network information, and the event information according to the requirements of users for a particular type of news and information to track , then timely report the certain public opinion for the widely attention.

2.1. The Main Structure

System includes network data fetching part, fetching the data processing part, and analyzes the processed data part and display part of the public opinion analysis results. The network data fetching part is fetching some concerns with the Internet information, including news, BBS, blog, micro blog [9], etc.; Fetching the data processing part is to catch the web page for further cleaning and discarding the useless, to retain only useful information; The data analysis part, through the data of cleaned of steps, such as Chinese word segmentation by methods such as classification, clustering to obtain information in the public opinion hot: Public opinion analysis shows the hot public opinion after the part is the analysis results in different ways. The main structure shown in Figure 1:

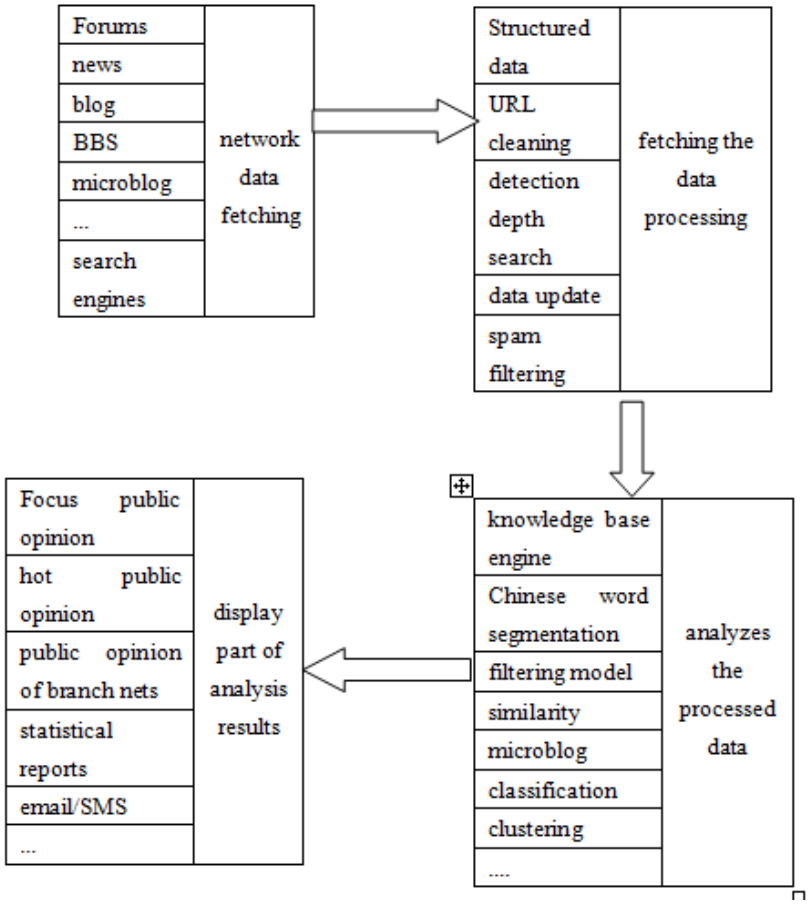


Figure 1. Network public opinion analysis system structure diagram

2.2. The Web Crawler

The first step of public opinion analysis work is to collect relevant information, mainly through the search engine system. Main parts include: document extraction subsystem, document filtering subsystem, document processing subsystem, the index retrieval subsystem and output subsystem. Shown in Figure 2:

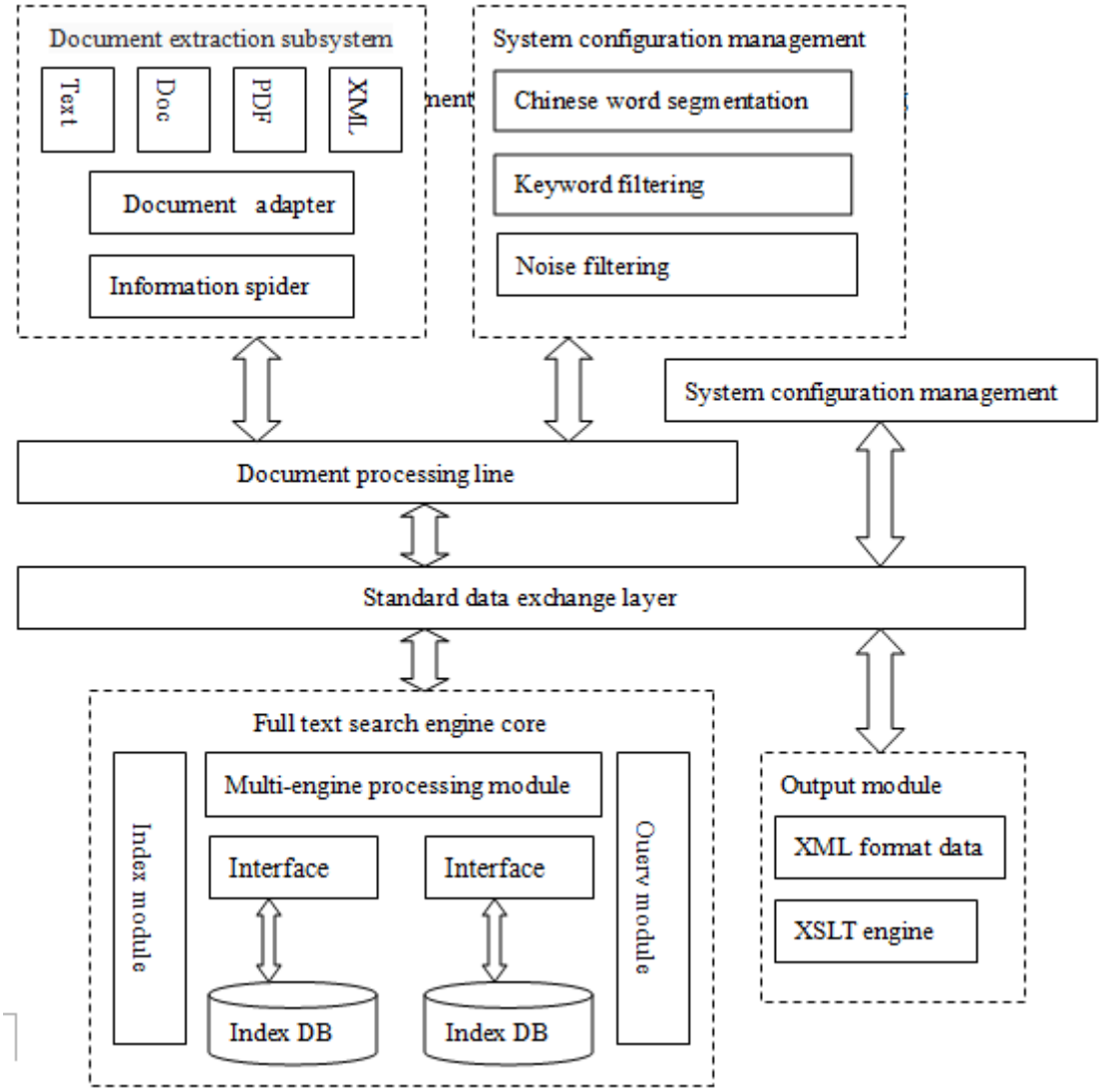


Figure 2. Network public opinion analysis system structure diagram

One of the most important is the document extraction subsystem, also called web crawlers, mainly composed of document adapter and the information spider. Document adapter is used to deal with different types of documents; the information spider is mainly responsible for the page information collection work. Document extraction subsystem, according to the provisions of the configuration file, first time to produce the spider on the distribution of information on the Internet information node traversal scan type, and then call the corresponding document adapter extracting network document information. Document the adapter can extract all kinds of page file.

There are many open source web crawler, such as Wget, Httract, Larbin etc. Their function is similar, the main difference is in the performance, but the main factors influencing the performance is web spiders crawling after storage, this paper uses a Larbin web crawler, its structure shown in Figure 3:

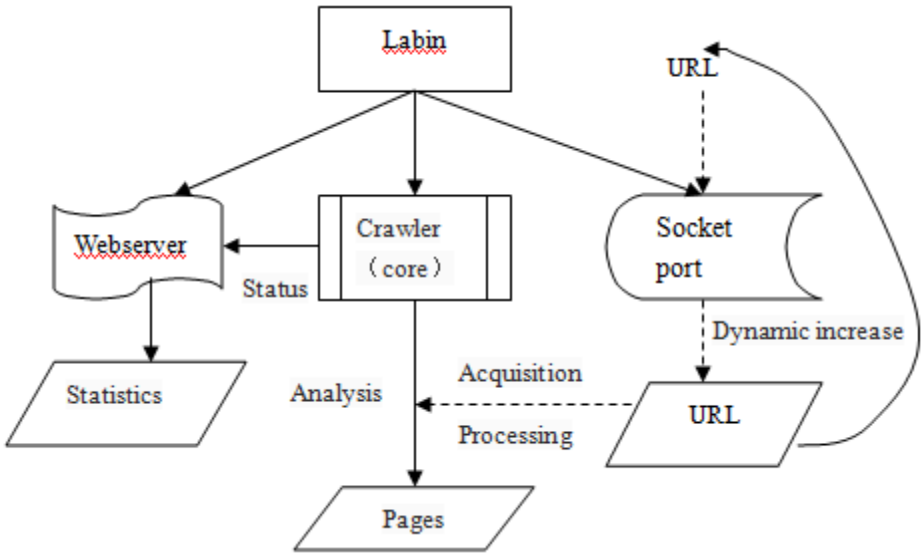


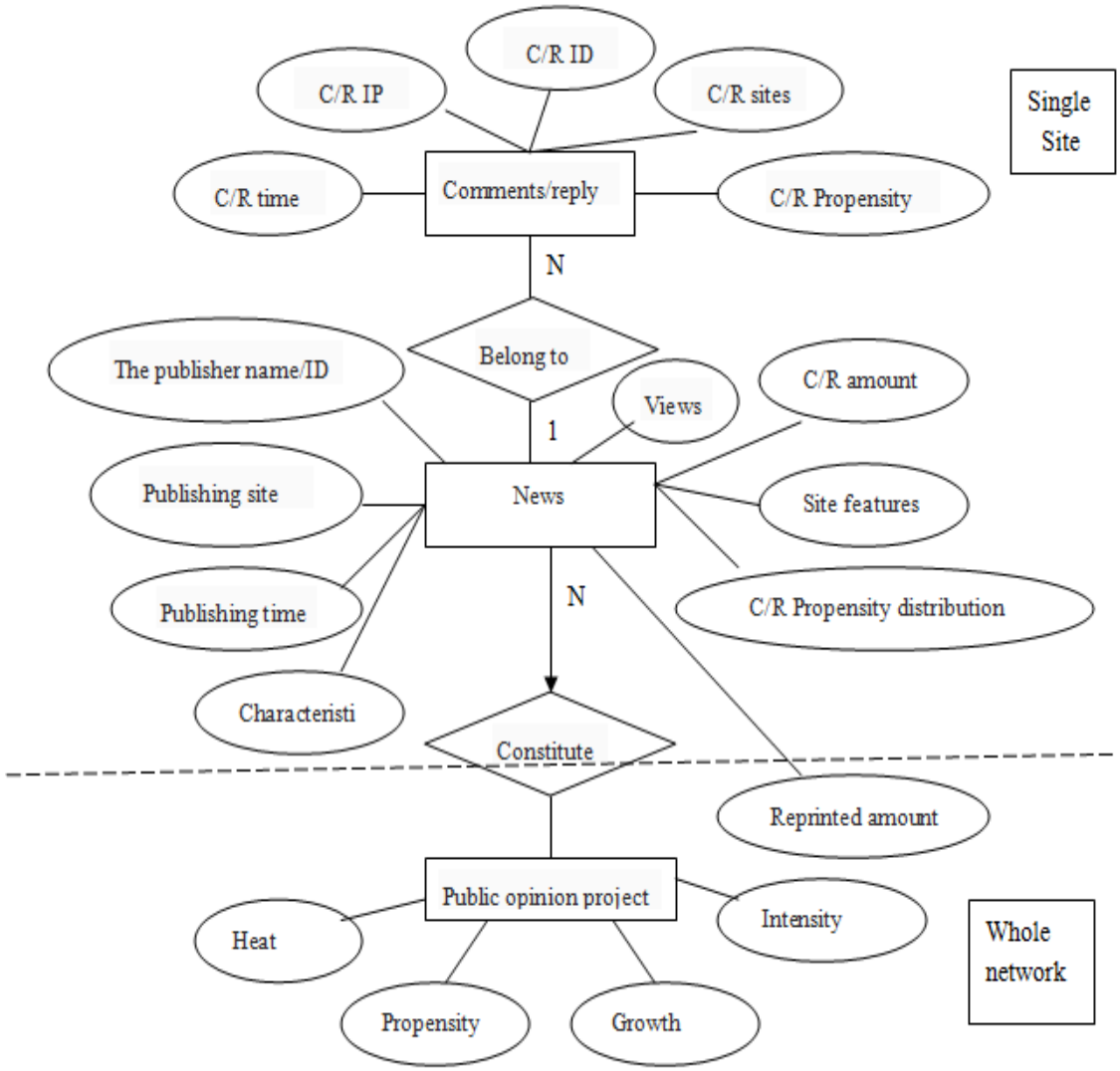
Figure 3. Labin crawlers structure

Larbin contains three parts: the crawler, check the state of the crawler webserver and used to receive a new one URL socket port. The crawler is the core of larbin function, acquisition, analysis, processing the URL to get the page; Webserver for viewing the crawler to download the current state, given statistics; Socket port used to receive the URL in the crawler running dynamic increase need to download the URL.

2.3. E - R Analysis of Public Opinion System

Public opinion analysis system needs to be based on network data capturing and protocol reduction as well as the web crawler to get the data, and to obtain the data content of extraction, extraction after segmentation, classification, and thus to obtain hot words, post count/reply to browse and participate in discussions staff analysis of the characteristics of a sensitive topic, etc., coupled with the secondary search tools and public opinion report output auxiliary tool campus BBS public opinion. E - R analysis on the public opinion of the whole system can get all the entities in the public opinion system and related properties, heat transmission of network public opinion, the strength of content, the audience orientation and growth rule is the most important properties of network public opinion special characteristics.

System analysis was carried out on the factors of public opinion, can draw a E-R diagram of the system, as shown in figure 4:



3. Improvement of Single - Pass Clustering Algorithm

Public opinion analysis module is the core of the network public opinion analysis system, mainly includes: subject identification module, the words Topic tracking module, subject evaluation module [10]. The realization of a new topic used a Single-Pass improved clustering algorithm; Multicenter forms can reflect the variation of the public opinion topic; Use a double or multiple keywords more gives higher weights method can accurately identify the subject. The module design idea is:

Using multicenter first used the vector and the amount of text vector to two-way comparison, comparison principle is adopted in the process of double or multiple keywords to give more power value, and make the topics into topic clustering tree.

Divided into the parent class topic is: read all document title vector and then compare the similarity.

Sub classing topic is: reading within class document text vector and then compare the similarity.

Basic idea is: first of all, we will be the title of the document feature vector and the extraction of text eigenvector effectively; Individual characteristic vector clustering if according to the document title vector as a standard, can be divided into the first hierarchy clustering, a document title vector are compared with those of similarity of the parent class topics can be divided into class a topic; If a document could not be determined document title vector comparison, will the title text vector and to compare the similarity of the parent class topic, in order to determine the class of topics, this is the second partition clustering.

Improvement of Single - Pass clustering algorithm process is as follows:

Start:

(1) The initial loading for conversation class;

(2) Di read the document

(3) compared with the parent class topic title feature vector similarity ;

(4) Determine whether more than threshold DC1?

Yes: step (5);

No: text feature vector similarity compared to the parent class topics, determine whether more than threshold DC1?

Yes: step (5);

No: logo for a new parent class topic, go to step (6);

(5) The classification to the corresponding parent class topics, are compared with those of text feature vector similarity subclass topic, determine whether more than threshold DC2 again?

Yes: the classification to the corresponding subclass topic, step (6);

No: logo for a new subclass topic;

(6) To judge all data have been processed?

Yes: update and store the data, end the algorithm;

No: to continue with the next document Di + 1, go to step (2).

4. The System Implementation and Experiment

The test of this system in a university network center, which uses the Mysql database, operating system of Red Hat Linux Enterprise 4 0. Test data is line of 5.5 million public opinion information data; raw data size is 2 g

The experimental steps: index cache read records in the database, and storing the data in the index, and then optimized. If there is no set cache is direct cycle query 20 times, and then calculated the average value of the query time. Repeat the above process 10 times, calculate the average time. 50 records before finally returned to the data, and descending order according to the relevant work.

The experimental results:

Index and the optimization time were 9000 seconds, form the index file size is 3.22 GB. To retrieve a single word, the size of the test result set for 100000, 200000, 300000, 800000, 1 million, 2 million, 4 million, 5 million, when the performance. If the result set is less than 400000, Lucene.net speed is faster, but the retrieval time increasing with the increase of the result set is also slow. The results as shown in table 1:

Table 1. Search results

Word	Results	Time
Internet public opinion	901241	80.4
Computer network	2593142	73.2
Entertainment	5196317	109.6
employment rate	897625	99.8

Finally, System in this paper and the typical general search engines such as baidu, sogou, Google, etc. The experimental comparison, mainly in the field of Internet public opinion information searches for the object, the results are as follows:

Table 2. Contrast this system with independent search engine

	Our system	Google	Baidu	Sogou
Number of results pages	896	601	641	759
precision	0.91	0.78	0.82	0.66

Contrast table above shows that on the premise of random term public opinion, public opinion in terms of recall ratio of search results information search engine medicine for about a

quarter higher than general search engine. While precision is about 21% higher than general search engines. This is embodies the superiority of this system.

5. Conclusion

Network public opinion is the political beliefs of the people through the Internet in a variety of government and social phenomena, the problem expressed by the sum of the attitudes, opinions, emotions, with freedom and concealment, interactivity and timeliness, rich diversity and other characteristics. University network with a common network public opinion public opinion in common, but also has its own peculiarities. On hot topics, emergencies, major issues for quick identification and tracking, can help students understand the school the focus topic, you can boot the campus public hotspots in quickly grasp and manage information in terms of public opinion, the campus network, for schools an effective channel [11]. This article will examine the combination of theory and empirical research, proposed information collection model network of public opinion through search engines and web crawlers' combination of the establishment of the network public opinion collection system to meet the different needs of people for information collection. Finally, the demand from the network information collection and analysis of public opinion starting to study public opinion information network analysis system, realized from the URL crawled pages to re-work, to analyze the information obtained to complete the campus network public opinion analysis system design.

Acknowledgment

This work is financially supported by Scientific research project of shaanxi province department of education, NO. 15JK2133.

References

[1] Yue Xiang Fen. Cluster Analysis on Internet Public Opinion Literature. *Pioneering With Science & Technology Monthly.* 2012; 8(6): 96-100.
[2] Sun HaoJun, Shan Guang Hui, and Gao Yu Long. Algorithm for high-dimensional categorical data weighted subspace clustering. *Computer Engineering and Applications*. 2014; 50(23): 131-135.
[3] Wen Shun, Zhao Jie Yu, and Zhu Shao Jun. Hierarchical Clustering Based on a Bayesian Harmony Measure. *Pattern Recognition and Artificial Intelligence (PR&AI)*. 2013; 26(12): 1161-1168.
[4] LV Gang, Chen Sheng-bing. An Improved Entity Similarity Measurement Method. *TELKOMNIKA Telecommunication, Computing, Electronics and Control.* 2014; 12(4): 1017 –1022
[5] Chen Liping. Design and Implementation of Data Collection and Extraction System on Public Opinion in Campus BBS. *Huazhong University of Science and Technology.* Wuhan, Hubei PR China, 2012; 45(6): 15-19.
[6] Manaev Oleg, Manayeva Natalie, Yuran Dzmitry. The'spiral of silence'in election campaigns in a post-Communist society. *International Journal of Market Research.* 2011; 26(53): 319-338.
[7] Yin Yantai. The Research on Development and Guidance of College Students' Network Public Opinion in China. *Hunan University.* 2013; (4): 11-17.
[8] M Ikonomakis, S Kotsiantis, V Tampakas. Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers.* 2010; 4(8): 966-974.
[9] G Guo hao. Research and Design of Public Opinions Information Search System Base on LUCENE. *PLA Information Engineering University.* 2012; (12): 21-24.
[10] Wang Qing, Cheng Ying, Chao Naieng. On the Construction of Internet Public Opinion Index System forts monitoring and Early Warning. Books intelligence work. 2011; 55(8): 55-58.
[11] S Siti Maimunah, Husni S Sastramihardja. CT-FC: more Comprehensive Traversal Focused Crawler. *TELKOMNIKA Telecommunication, Computing, Electronics and Control*. 2012; 10(1): 189 –191.