Key Techniques of Public Opinion Mining Based on Big Data

AN Dezhi[1,2, a*] WU Guangli[1,2], LU Jun[1,2], ZHANG Shengcai[1,2]and LI Yan[1,2]

1 School of Information Engineering
Gansu Institute of Political Science and Law, Lanzhou 730070, China
2 Key Laboratory of Evidence Science of Gansu Province
Gansu Institute of Political Science and Law, Lanzhou 730070, China
adz6199@gsli.edu.cn

***Abstract* — With the rapid development of technology, social online networks have become the platform to express public opinion, to discuss public affairs, to participate in economic, social and political life. It also brings public opinion of relevant government departments to become network management issues. How can public opinion hot issues be detected electronically? Correctly guiding public opinion trends is currently needed urgently to solve the difficulties. This paper analyzes the big data era of network status and characteristics of public opinion, through the study of the concept and the main characteristic of big data technology. It explore how big data in network public opinion works, the provision of technical information for the network of public opinion to resolve reference work, as well as methodological support.**

Keywords -- Big Data; Public Opinion Mining; Data Mining; Web Crawler

I. INTRODUCTION

Public affairs material network of media attention, on the one hand from traditional media such as newspapers, radio and television coverage, on the other hand from the original network, especially on a large BBS users self-publish some news and rumors. If the traditional media coverage reproduced by the portal, and be prominent advocate a thematic or Internet rumors get confirmed by the traditional media, online media view has been the traditional solidarity, so that online and offline interaction, will lead public opinion raging, create a strong wave of public opinion, especially the parties bear responsibility for the social management of public authority formed pressure, momentum will control the executive branch of great impact. Therefore, the current relevant administrative departments of public opinion for network control becomes particularly critical [1].

At present, social conditions and public Internet delivery is becoming an important basis for the executive branch decision-making. However, the information behind massive Internet, but also to hide some of the content of yellow, violence, and the spread of speed is often beyond people's imagination [2]. How to eliminate "content threats" such harmful information, how to extract the valuable public opinion in so many public opinion in a timely and effective to provide a useful reference for the work of government departments to highlight the very important and relevant government departments Can the first time the probe and control public opinion trend is particularly critical.

Big Data is not only new technologies but also methodology, focusing on large data mining valuable information that will be applied to large data network public opinion work, it is a new requirement under the new situation and new environment to deal with public emergencies. Mining can be hot issues and focus areas are concentrated in the website information, and through professional techniques based on information collected large public opinion gradually filtering and pre-processing of data, these vast amounts of information in a visual way to show it to quickly and accurately, intuitive understanding of the various movements on the network.

II. CHARACTERISTICS OF PUBLIC OPINION UNDER THE BIG DATA

As the most concentrated area from the media, social networking platform has been a key area of public emergencies network monitoring public opinion, and in recent years the popularity of the Internet in rural areas has been successful, the gap between urban and rural areas gradually narrow the popularity of the Internet, the network is no longer the right to speak confined to the city's young people, a number of Internet users in rural areas and the urban fringe to start a direct reflection of feelings of public opinion and social issues through the network to express their understanding and perception of events, online variety of real or false, rational, non- rational speech are superimposed, confuse, to the Internet and network supervision of public opinion monitoring challenges [3-4]. Compared with the past, the traditional media, mass media, network media era of big data mainly has the following characteristics:

1. Relatively open Public Information. Social network service (SNS) so users grasp more and more right to speak directly to users in the state of nature to express personal emotions, speech point of view, a variety of opinions from all sides to break the pattern of a unified voice of the mainstream media, network media into free open from media age.

2. The Public Dissemination of Information Quickly. When public emergencies, can the social network by phone "broke", the event spread quickly in the event of Internet

users in the network field occurs from the event to spread the dissemination of information generated from the collection of individual opinions to community comments, advice from a comprehensive to the formation of public opinion, in the era of big data, almost a "nuclear fission" type of "butterfly effect" process.

3. Public Information Rich and Varied. In the era of big data, network information is vast and diverse. In content, network media inclusive, both positive and healthy speech, but also emotional expression of speech and personalized irrational; propagation form, podcasting and microblogging combining audio and video social networking applications have sprung up everywhere, microblogging, micro letter and other social networking tools are no longer limited to plain text AC applications, but integration of photographs, video, network intercom, podcasts and other forms of multimedia audio and video network applications, the network in the dissemination of public information more diversified forms.

4. Tendentious Public Information. Since the time of the incident, the incident scene mobile phone users affected some emotional factors, knowledge of the event, the presence of love and hate speech perception and preferences propensity catch afraid photographs may not be the full picture, the event itself and some sources of information credibility is not high, but some irrational layers of network remarks were forwarded, spread rapidly, and some have even been processed, turned into a rumor, to the government response to emergencies caused great passive and disposal.

III. PUBLIC OPINION AWARE BASED ON BIG DATA

Network is primarily a sampling survey method, the result accuracy with the increase of sampling randomness, relationship with the increase of sample size is not big, that is to say, the randomness of the sample is more important than the number of samples, but such randomness is very difficult, so much so that if the sampling object is Internet users so that complex and massive object, it is hard to find a standard, the optimal sampling is less likely to want to get small samples can accurately reflect all the characteristics of the whole [5].

Big data is not only refers to the mass of data, but including massive amounts of data and massive data processing method. Internet public opinion is not a direct presence in the online world of data, but by the relevant technical after extracting and analyzing the results come from a massive network data. Online Public Opinion get just reflects the thinking of big data. Big Data technology model is simple, good scalability and fault tolerance and high parallelism is an effective tool for dealing with the network of public opinion. Identification of the basic network of public opinion based on the large data flow shown in Figure 1. By the web crawler module from the preset network forum to get network information document, and in fact, when stored in a database, clustering and classification.

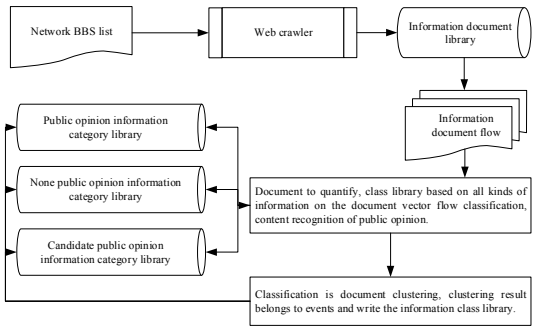

Figure 1. Process of the public opinion recognition based on big data.

1. Using web crawler tool in the Internet page document information in real-time collection, collection of web pages documents including news, forums, web documents and blog Web documents, since the page document contains many HTML tags, as well as announcements, navigation and other irrelevant information, so information extraction unit by parsing pages pretreatment acquired title, text and other information assembled into a web page information document and save it to the database.

2. Export page information in the database to document and quantify the document establishing public opinion class libraries, non-public opinion public opinion class libraries and class libraries candidate to hold the information about the document type information of public opinion, and based on the event sub category of each information repository class category of the document vector traffic classification, classification of information written to the corresponding class library.

3. For the classification step in the document cannot be identified as a known type of document vector of public opinion, by clustering to identify where there may be a new event or a new public opinion, public opinion category. The number of documents obtained poly result clustering if the cluster contains more than a given threshold, the establishment of the cluster classification model.

IV. WEB CRAWLING TECHNOLOGY ON PUBLIC OPINION INFORMATION.

Web public information retrieval the main advantage of the page description information, such as page title, content, links, etc., through the analysis and processing of such information eventually enter keywords approximation based on the results to the user, according to the priority from high low returns to the user link. However, information retrieval for Web space, consider the similarity note page text only enough space picture page contains links to spatial data reflect the real number of implicit dig spatial information, so the Web pages Get in touch with retrieval of spatial data is public data on Web crawling be an important part of the Institute to be considered [6]. Web crawler is an automated program crawling web content, which by downloading Web pages, while drill track pages include hyperlinks page to find useful information, or until certain conditions are met cannot

find new links, operating mechanism as shown in Figure 2. A search engine performance and scalability are often affected by the processing power of web crawlers.

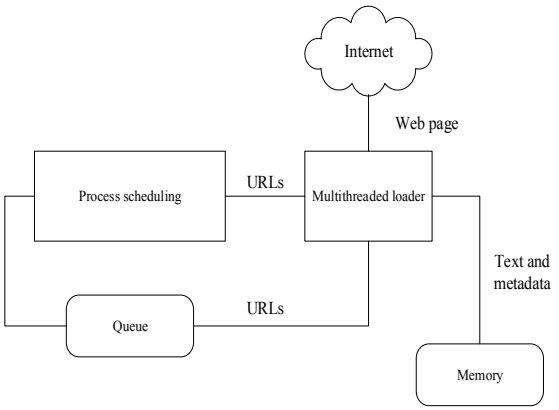

Figure 2.The running mechanism of Web crawler.

Ordinary reptile reptiles after starting operation strategy is its analytic seed URL, Get Link URLs, and extract its title and keywords stored in the local database, the original database has been recorded link address, check whether the data update (based web date) If you have updated and re-extract content main keyword, or not treated; has a new link is added to the database. Because of the diversity of network information, there are differences in their crawling strategies, such as frequent changes of certain content (news, Blog, etc.) are generally shorter intervals crawling once and for static pages, then crawled a few months.

V. Overall Scheme based on Public Opinion Mining based on Big Data

Public opinion is based on data mining on the Web for the study, to dig useful knowledge, and knowledge for effective management, to provide users with intelligent services. The main function of public opinion is to provide web information mining excavation, analysis, retrieval functions. The system of text messages and other relevant information are processed separately parsed to establish the concept of semantic indexing text information for user retrieval.

Data mining is a long-term research and application of database technology is the inevitable result of the development of database technology, but a more advanced stage, it cannot only large amounts of historical data query, data can also be found in the history of the unknown potential link [7-8]. Faced with the current status of large amounts of data, and association rules is an important branch of data mining is to study an advanced and intelligent data processing and analysis technology has become a hot spot. By association rule mining, a lot of useful information can be implied in the sea there is the potential value of the data [9]. Target association rules is an effective method to extract the most interesting patterns. So far, it has been proposed many effective association rule mining algorithm, the algorithm proposed mining algorithm Agawal algorithm is based on a priori the most famous, but it is a challenge in the face of time and spatial scales, so many researchers to explore new mining methods, expanding the concept and application of association rules.

Data mining is a lot of data warehouse, reveal implicit, previously unknown, potentially important in the process of valuable information, data mining is a decision support process, which is mainly based on artificial intelligence, machine learning, pattern recognition , statistical, database, visualization technology, highly automated analysis of enterprise data to make inductive reasoning, dig out potential models to help users adjust the marketing strategy to reduce the risk, make the right decisions. From data mining process, as shown in Figure 3.

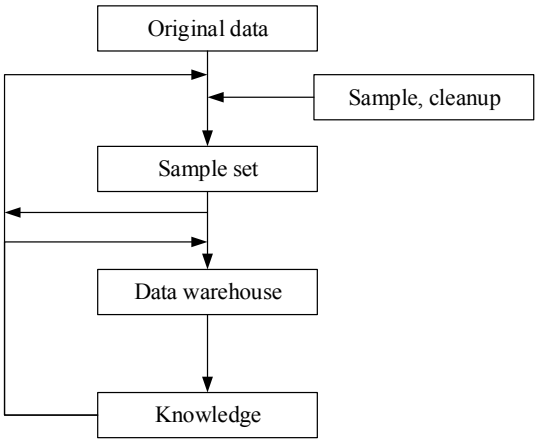

Figure 3.The basic process of data mining.

After data collection, sampling and cleanup needs. Cleanup is the result of sample data sets. A data warehouse is an effective data storage, data mining is very helpful. You can use different data mining algorithms. Sometimes, the need to return to the final stage of the above process.

Based on public opinion overall function of large data mining framework is built on the basis of CSP comprehensive web database mining policy from the Web on the acquisition and preliminary classification, the main function is to retrieve information for users to conceptualize and retrieved by the concept of information from the mass users really want information. The system consists of user management module, mining policy management module, mining module, fault management modules, mainly to complete the text analysis, text classification, conceptual clustering, semantic indexing, automatic summarization, event handling and other tasks. We formulate public opinion sources, including e-mail, Xinhua News Agency, Internet, BBS, blog and so on, through public opinion

Mining module to tap into the system servers, and text classification, clustering, and similarity of qualitative analysis, the final result after the selection of the sort output to the user interface, the specific process is shown in Figure 4.

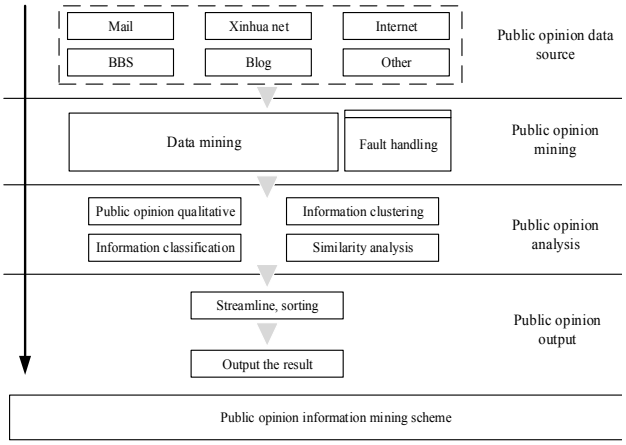

Figure 4.Public opinion information mining scheme based on big data.

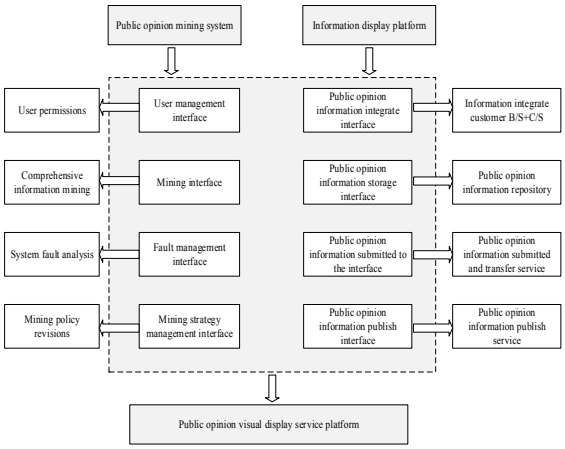

Figure 5. The public opinion information mining architecture based on big data.

VI. THE PUBLIC OPINION INFORMATION MINING ARCHITECTURE DESIGN BASED ON BIG DATA

According to user needs the ability to tap public opinion, combined with careful planning platform overall system design, the system targeted at the intelligence information system integrated application platform integrated mining judge platform, and is a judge of digging tool system. As a comprehensive judgment of intelligence information tools, the role of this system is that mining data from public opinion, public opinion designated repository accessible through a series of analysis to determine the model of public opinion generated judging intelligence and clues type of information available to the intelligence Applications department and command decision makers. To this end, we in-depth analysis and a deep understanding of user needs, using the advanced nature of mainstream technology, practicality to the concept of user experience design refinement functions, data integration and application integration as the core of the comprehensive development of innovation The overall design of the route, to achieve the product of public opinion mining system. Therefore, the core function of this system is to tap public opinion, and output information comprehensive analysis and comprehensive analysis of the results of public opinion public opinion information. Public opinion based on big data mining architecture is shown in Figure 5.

VII. CONCLUSION

Compared to traditional information processing technology, big data technology has advantages of massive data processing, gradually brought for review by scholars. Big data is not just an information technology, but also a methodology, it has brought a new revolution in information technology, it represents a new era. From the features and applications to target large data, not just in the pursuit of large data information capacity and processing speed, but more important is the value of information, the large public opinion data actively on the network, under the new situation and new environment, new conditions suddenly demand response to public events. This paper outlined the basis of common network public opinion analysis methods, the proposed method of network public opinion analysis. To view the direction of big data analysis, we studied the use of large data mining and analysis techniques to explore the fragmentation of public opinion information integration process and public opinion model construction method. Network public opinion analysis electronically is beyond the existing common analytical framework, but we must be innovative in the use of big data to advance the technology.

ACKNOWLEDGEMENTS

The national natural science foundation of China, Research of the key technology of large-scale depth calculation group social dimension under the Web environment sensitive (61363024).

[1] Cambria E, Schuller B, Xia Y, et al. “New avenues in opinion mining and sentiment analysis”. IEEE Intelligent Systems, 2, pp. 15-21, 2013.

[2] Liu B, Zhang L. “A survey of opinion mining and sentiment analysis, Mining text data”. Springer US, 14, pp. 415-463, 2012.

[3] Liu B. “Sentiment analysis and opinion mining”. Synthesis Lectures on Human Language Technologies, 5, pp. 1-167, 2012.

[4] Pak A, Paroubek P. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, LREC, 10, pp.1320-1326, 2010.

[5] Culotta A. "Towards detecting influenza epidemics by analyzing Twitter messages", Proceedings of the first workshop on social media analytics. ACM, 13, pp. 115-122, 2010.

[6] Chen H, Zimbra D. "AI and opinion mining. Intelligent Systems", IEEE, 25, pp. 74-80, 2010.

[7] Sobkowicz P, Kaschesky M, Bouchard G. "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web". Government Information Quarterly, 29, pp.470-479, 2012.

[8] Murray G R, Riley C, Scime A. "Pre-election polling: Identifying likely voters using iterative expert data mining". Public Opinion Quarterly, 73, pp.159-171, 2009.

[9] Segev E, Baram-Tsabari A. "Seeking science information online: Data mining Google to better understand the roles of the media and the education system". Public Understanding of Science, 21, pp.813-829, 2012.