Research on the Application of Web Mining Technology in Early Warning of Internet Public Opinion

Longjian Lin, Guanjin Li

School of Information, Guangdong Polytechnic of Science and Trade, Qingyuan, Guangdong, 511500, China

Keywords: Web mining technology; Internet public opinion early warning; Emergency

Abstract: The diversity of network information determines the diversity of mining tasks. According to the different processing objects, it can be divided into text mining and multimedia document mining. Internet public opinion can not only develop into mass incidents, which will further lead to social crisis, but also the emotionalization of Internet public opinion will lead to the vicious development of various emergencies, which will also increase the difficulty of dealing with them. Web usage mining can also analyze and predict the behavior of netizens. These systems provide certain conditions and support for the analysis of online public opinion to a certain extent. Aiming at the outstanding problems existing in the existing internet public opinion early warning system, and according to the advantages of Web data mining in information analysis and knowledge discovery, this paper comprehensively applies Web mining, semantic analysis, information integration and other technologies, and constructs a internet public opinion early warning system model based on Web mining, so as to realize the automation, intelligence and real-time collection, analysis and processing of emergency internet public opinion and crisis early warning.

1. Introduction

In recent years, with the advancement of China's reform and opening-up and social transformation, the topic of emergencies and public crises on the Internet has been highlighted, and the Internet has gradually replaced traditional media as a new field of public opinion, and the ecological environment of social public opinion based on the Internet has gradually formed [1-2]. Internet public opinion can not only develop into mass incidents, which will further lead to social crisis, but also the emotionalization of Internet public opinion will lead to the vicious development of various emergencies, which will also increase the difficulty of dealing with them. Web mining, as a new information mining technology, can effectively obtain and analyze related public opinions from the Internet, achieve the purpose of early warning and decision-making, and provide great help for early warning of online public opinions. This paper will discuss the basic concept, technology and application of Web mining.

2. Web mining technology

Web information capture and so on. The diversity of network information determines the diversity of mining tasks. According to the different processing objects, it can be divided into text mining and multimedia document mining. The association in data can be divided into simple association, time series association and causal association, etc., which are usually mined by establishing association rule base for comparison and analyzing the covariation among attributes [3-4]. By using correlation analysis, we can find out the relationships among various public opinions on the Internet in the same period, and provide help for comprehensive analysis of online public opinions.

The basic goal of data mining technology is description and prediction. By depicting the potential patterns in massive data, and forecasting according to the potential patterns in the data, valuable models and rules in the data can be found. The commonly used methods of data analysis by data mining mainly include classification, regression analysis, clustering, association rules, characteristics, change and deviation analysis, Web page mining, etc. [5-6], which mine data from

different angles. In order to find the difference between the observation result and the reference quantity, the deviation of the observation result from the expected value is observed by the change and deviation analysis method. By using Web pages to mine the massive data of the Internet and collect all kinds of Internet information, we can find out what topics on the Internet are receiving attention and mine the public opinion information needed by the government, enterprises and social organizations.

In order to obtain and analyze the public opinion data on the Internet, we need the help of a network information acquisition tool-web crawler. Web crawler can automatically download web pages from the Internet and extract useful information from them. It adopts multi-thread concurrent search technology, and constantly crawls automatically among the nodes of the Internet [7]. Most of the captured network information data are stored locally in text or XML format for analysis. The work flow of the web crawler is shown in Figure 1:

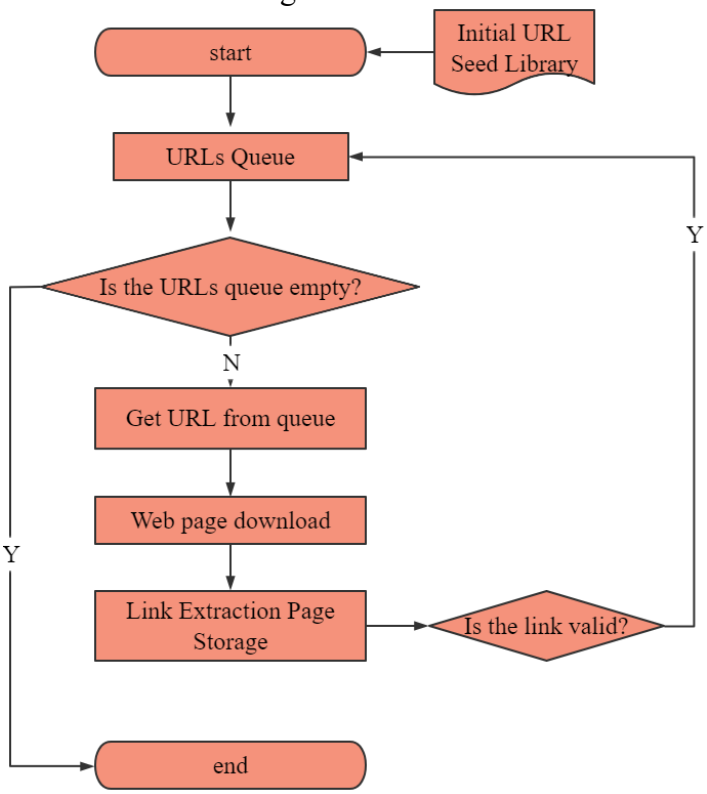

Figure 1 Work flow of Web crawler

At present, the content of online news reports is all original text content, which is unstructured. It is not possible to directly process the text input algorithm, but it is necessary to transform the text into information that is easier to be recognized by computers, that is, to formalize the text and get the result of formalization [8-9]. Vector space model is a document representation model based on word frequency statistics. Its basic idea is to use vectors in vector space to represent texts, and the similarity between texts is measured by cosine of included angle between vectors.

If the two vectors d_i, d_j are respectively represented as $d_i = (w_{i1}, \cdots, w_{in}), d_j = (w_{j1}, \cdots, w_{jn})$. The degree of similarity or correlation between two vectors d_i, d_j is usually measured by the included angle cosine $Sim(d_i, d_j)$. The cosine of the included angle of the vector is shown in formula (1):

$$Sim(d_i, d_j) = \frac{\vec{d}_i * \vec{d}_j}{|\vec{d}_i| * |\vec{d}_j|} = \frac{\sum_{k=1}^{n} w_{k,i} * w_{k,j}}{\sqrt{\sum_{k=1}^{n} w_{k,i}^2} * \sqrt{\sum_{k=1}^{n} w_{k,j}^2}} \quad (1)$$

If the two vectors have similar weight distributions, the smaller the value of $Sim(d_i, d_j)$,

indicating that the two vectors have high similarity.

Web mining mainly refers to the application of data mining on the Web, which comprehensively uses intelligent technologies such as data mining and natural language processing to extract content that people are interested in. According to different mining objects, Web mining can be divided into three forms: Web content mining, structure mining and usage mining. Web usage mining is to better provide intelligent services for users by mining users' online information and other resources [10].

In addition, Web usage mining can also analyze and predict the behavior of netizens. To a certain extent, these systems provide certain conditions and support for internet public opinion analysis. However, on the whole, the functions of these softwares have not reached the level of intelligent internet public opinion analysis, and all of them have some shortcomings. At present, a complete system has not been formed.

3. Early warning mode of internet public opinion based on Web mining technology

3.1. Text mining

Information extraction technology tries to extract specific words and the relationship between specific words from the text, so as to discover new and meaningful information and knowledge. This kind of technology is the basic point for a large number of text mining algorithms. After the information is extracted, the vector space model is used to represent the content of the document or text [11].

The mathematical expression of TF-IDF is as follows:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \quad (2)$$

tf_{ij} is the number of occurrences of feature items in the text, w_{ij} represents the weight, and N and n represent the total number of texts in the training set and the number of documents containing feature items, respectively.

By using the method of text mining, we can monitor the Internet information, extract the characteristic words, classify the characteristic words, make statistical analysis on the topic opinions and tendencies, and speculate the trend of Internet public opinion. Figure 2 shows the text data mining method.

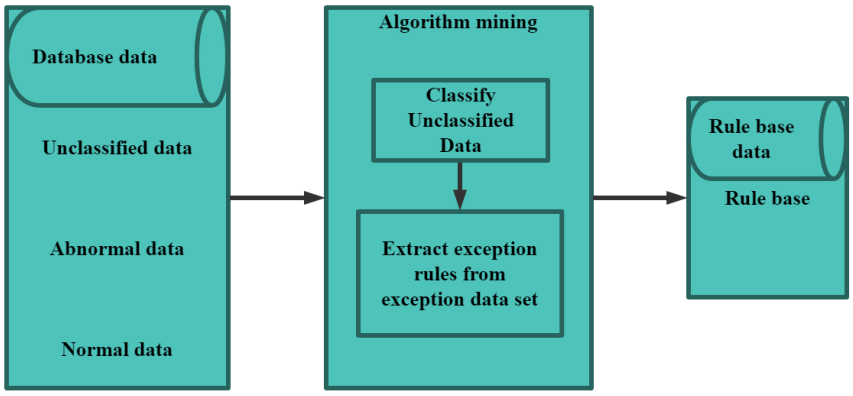

Figure 2 Text mining

There are also difficulties in the application of public opinion monitoring technology based on text mining in analyzing users' emotional tendencies. How to use data mining technology to analyze the emotion of these online texts is a hot issue in public opinion monitoring and research at present. However, at present, emotion analysis of user-generated content is inaccurate, so it is necessary to further study it.

3.2. Hot spot public opinion discovery

The difference between Web mining data and traditional data or data warehouse is that the target

information is dynamic, confusing and basically unstructured or semi-structured. It is impossible to be successful and effective to adopt direct mining technology like traditional data and data warehouse, so Web mining must have necessary data processing links. When a report discusses the events or activities involved in a topic, it can be considered as belonging to the topic. The purpose of topic discovery is to gather different reports on the same topic scattered on various websites under the same topic and describe a topic in a characteristic way.

Because of the openness and anonymity of the Internet, netizens can often express their truest opinions and feelings in their comments. The aspects with obvious opinions are usually more interesting to netizens, which can better reflect the theme and focus of the news report. Moreover, the weight of feature words in the report vector can be adjusted through the quantitative expression of opinion tendency and strength of feature words, so as to obtain more accurate feature word weight.

The purpose of establishing the prediction model of public opinion development trend is to quantitatively predict the development trend of online public opinion, and to assist decision makers or public opinion supervision departments to grasp the trend of hot topics as soon as possible and intervene in time. Through these indicators, the development trend of hot topics can be expressed quantitatively.

The process of establishing a public opinion prediction model is to pre-process the original historical data of public opinion to get the pre-processing results suitable for algorithm analysis, and then establish a public opinion prediction model through an appropriate prediction algorithm. After the model is established, enter the implementation and analysis stage of the model, and input real-time public opinion data. After the model is processed, the prediction results of public opinion development trend will be obtained.

3.3. Web mining-based internet public opinion early warning system for emergencies

Analysis and early warning of internet public opinion is a frontier field that integrates multi-disciplinary knowledge such as computer network, artificial intelligence, data mining, natural language processing, etc. It involves the whole process of internet public opinion information collection, analysis, processing, classification, monitoring and early warning. Web usage mining is a process of mining Web usage data or access logs to extract the behavior patterns of visitors and obtain valuable information. Web page access frequency and other knowledge patterns, so as to better understand user behavior and provide intelligent services. Through Web usage mining, we can determine the hot spots and focus of public opinion and predict the behavior of netizens. Through Web structure mining, we can obtain the semantic knowledge of links highly related to public opinion topics and the logical structure of links, thus helping public opinion analysts to determine important public opinion sources and central pages.

Web access data mining is to mine the access modes of users visiting Web sites, and the mining object is the log file records on the server, including Server Log Data. It is based on the user's recent access characteristics and monitoring the user's access behavior to get the user profile file. It is very important to cluster users according to their access data, put the mining data in the model, and analyze whether the calculated results are consistent with the real situation. Therefore, it is necessary to carry out experiments continuously until the accuracy of the results is higher than the original standard, so that the results are valid and the established model is correct.

This paper comprehensively applies Web mining, semantic analysis, information integration and other technologies to build a internet public opinion early warning system model based on Web mining, as shown in Figure 3.

This model includes three layers: public opinion collection layer, public opinion analysis layer and early warning application layer. It integrates and integrates the important functions of the whole process of public opinion early warning of emergencies, and realizes the automation, intelligence and real-time of public opinion collection, analysis and processing of emergencies and crisis early warning.

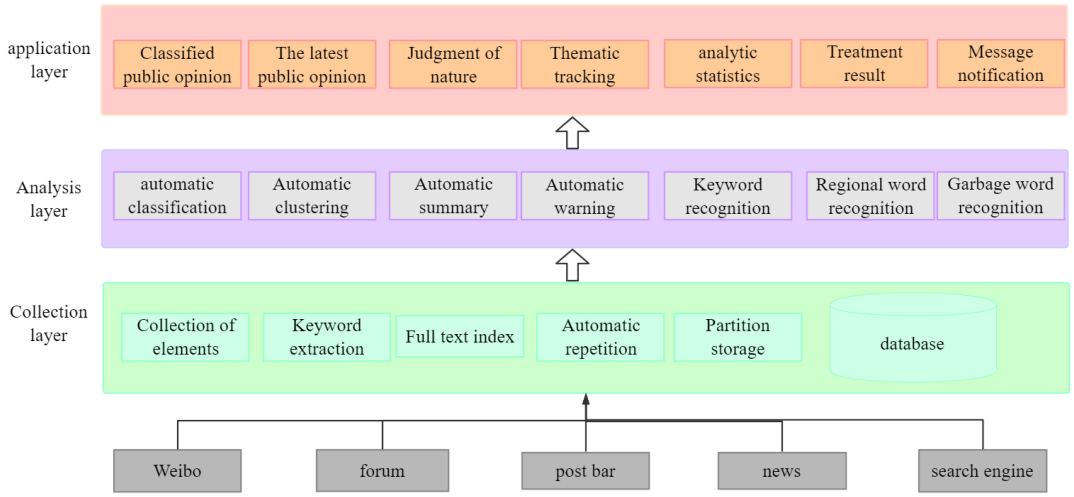

Figure 3 Web mining-based internet public opinion early warning system model for emergencies

The main function of the system framework is to provide a unified platform to support various applications to load, communicate and cooperate with subsystems, and to encapsulate the database and communication operations of various services, so as to provide a transparent channel for business subsystems to use. Each business subsystem is developed by using the interface specification published by the framework, that is, it can realize the coordinated operation of the whole system.

The system framework also includes the functions of basic modules such as user management and log management. Establish the tasks of automatic forecasting and early warning and correlation analysis, and set the scheduling strategy according to the needs, so that the system service program can automatically start the related tasks and realize the purpose of automatic forecasting and early warning.

4. Conclusions

The analysis and research of Web public opinion involves related knowledge in many fields, such as data mining, text mining, Chinese semantic analysis, Web information capture and so on. Web mining, as a new information mining technology, can effectively obtain and analyze related public opinions from the Internet, achieve the purpose of early warning and decision-making, and provide great help for early warning of online public opinions. Integrating Web mining into the analysis and early warning of internet public opinion in emergencies can give full play to the advantages of Web mining technology in processing massive network data and discovering hidden knowledge rules, realize automatic and intelligent acquisition of internet public opinion information and deep and multidimensional analysis, and achieve the purpose of dynamic early warning and auxiliary decision-making of emergency internet public opinion.

References

[1] Srinivasan, S., Pandian, P., & Senthil. (2016). A unified model for preprocessing and clustering technique for web usage mining. Journal of multiple-valued logic and soft computing, 26(3/5), 205-220.

[2] Panigrahi, R. , & Srivastava, P. R. (2018). Understanding the motivation in massive open online courses: a twitter mining perspective. International Journal of Web Based Communities, 14(3), 228-248.

[3] Guo, H. (2021). Research on web data mining based on topic crawler. Journal of web engineering(4), 20.

[4] Gupta, M. K. , Govil, M. C. , & Singh, G. (2018). Text-mining and pattern-matching based

prediction models for detecting vulnerable files in web applications. Journal of Web Engineering, 17(1-2), 32-48.

[5] Kotenko, I. , Chechulin, A. , & Komashinsky, D. (2017). Categorisation of web pages for protection against inappropriate content in the internet. International Journal of Internet Protocol Technology, 10(1), 61-71.

[6] Luo Ping,&Wu Bin. (2020). Network Public Opinion Big Data Communication Feature Mining System Based on Artificial Intelligence. Modern Electronic Technology, 43(4), 4.

[7] Li Zhenpeng, Chen Bizhen,&Luo Jingyu. (2020). Research on the Classification of Internet Public Opinion Based on Text Mining. Systems Science and Mathematics, 40(5), 14.

[8] Hao Asia, Zheng Qinghua, Chen Yanping,&Yan Caixia. (2016). Abnormal Behavior Recognition Based on Internet Public Opinion Data. Computer Research and Development, 53(3), 10.

[9] Cao Ruijuan, Jiang Rengui, Jie Jiancang, & Zhao Yong. (2020). public opinion monitoring and evolution mechanism of urban waterlogging network based on big data. Journal of Xi 'an University of Technology, 36(2), 8.

[10] Qiu Zeguo,&He Baiyan. (2021). Analysis of Internet Public Opinion Clustering and Emotional Evolution Based on pca-spectral-lda: A Weibo Text Mining Study. Systems Science and Mathematics, 41(10), 13.

[11] Tang Huimin Fan Hesheng. (2017). Ecological Governance of Internet Public Opinion and Government Responsibility. Journal of Shanghai Administration College, 018(002), 95-103.