Design and research of university Internet public opinion analysis system based on web spider

Pingping Jia [1,*], Xiaodong Guo [2]

[1] Student Work Department Shandong Women's University, Jinan, China

[2] School of data science and computer Shandong Women's University, Jinan, China

*Corresponding author e-mail: jiapingping@sdwu.edu.cn

Keywords: University Internet public opinion, Web Spiders, Topic detection, Emotional analysis.

Abstract: The university Internet public opinion analysis system starts from the retrieval of multi-source data, focuses on the public opinion of university teachers and students' speech, detects the hot topics and events concerned by them, and analyzes their views, attitudes and emotions on these topics and events, which has a practical value and a significance in exploratory for university public opinion management and guidance.

1. Introduction

With the rapid development of computer technologies and the diversification of Internet media, people begin to use social platforms such as micro-blogs, WeChat, online forums and Twitter to express themselves and convey ideas. Internet media has become the main channel for college students to obtain news information, and it is also the main media for information transmit. Internet public opinion is the sum of cognition, attitude, emotion, will, opinion, viewpoint and behavior tendency expressed by users in the face of hot events, specific topics, and social phenomena. Teachers and students in colleges and universities are active in thought and have a strong connection with the Internet. They like to pay attention to social hot spots and comment on events related to their own interests or campus emergencies. The impact of Internet public opinion cannot be ignored. The analysis of Internet public opinion in universities is becoming more and more important.

In this Internet era of information explosion, Internet public opinion analysis has gradually become a hot topic. Foreign countries began public opinion analysis research as early as the 1990s. The most famous is the TDT (Topic Detection and Tracking) project of the U.S. Defense Advanced Research Projects Agency. TDT is an information processing technology, which is designed to detect continuous news streaming media information topics and continuously track existing topics [1]. Domestic research on Internet public opinion analysis started late, but in recent years, with the rapid development of the Internet, more and more researchers participate in it. At present, most of the data sources of the Internet public opinion system in colleges and universities are BBS forums, micro-blogs or news information in the school. The data sources are relatively single, while the comments made by college teachers and students are more multi-source, which can be forums, post bars, micro-blogs and news. In addition, the current public opinion analysis methods are relatively simple or preliminary, and there are some shortcomings, such as low accuracy and imperfect function of detection methods. Based on this, the sources of Internet public opinion information in this study mainly include the official website of colleges and universities, various communities of college students, forums, post bars, micro-blogs, WeChat, etc., using web spider technology to obtain data, and improving the accuracy of text analysis by combining text analysis and natural language processing technology. It can provide guidance for the school to do a good job in public opinion management.

  DOI: 10.25236/iccem.2021.035

2. Web Spider Related Technologies

2.1 Web Spider

The basic principle of web spider is to search for information by constantly grabbing URLs. The basic process of web spider system consists of three parts. First, send the request. The web spider sends a request to the target website through the network protocol and waits for the response of the server. Second, access to information. This step is a very important for the web spider to facilitate subsequent data processing. After the server responds, the web page source code will be obtained. The data can be obtained by analyzing the source code by constructing regular expressions or page parsing library. Third, data storage. After obtaining the information, the data can be saved in the local or remote database for subsequent use. The storage format can be selected according to the actual situation. The general web spider architecture is shown in Figure 1.

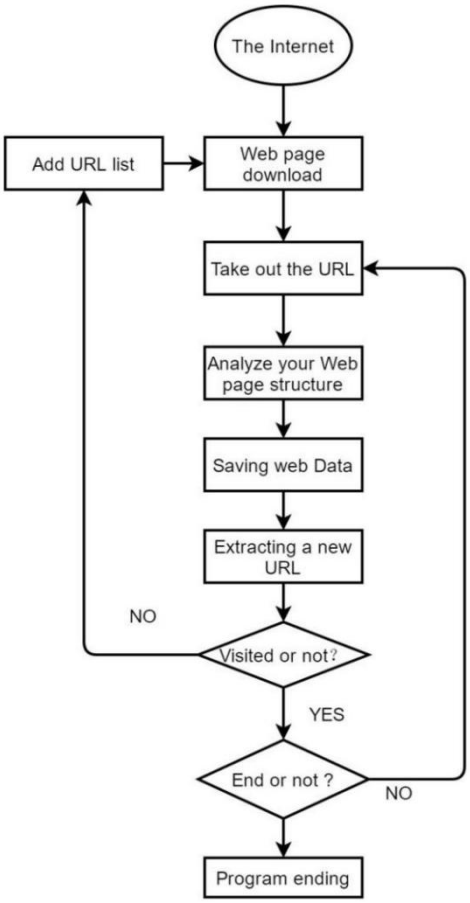

Fig. 1 Flow chart of a general web spider.

Scrapy is one of the most frequently used web spiders. It is a free and open-source framework to mine the data of the network or specific websites. Scrapy is an asynchronous processing framework based on Twisted and it is implemented in pure Python. It has a clear architecture, low coupling between modules, strong scalability and can flexibly meet various requirements.

A very important part of web spider is the crawling strategies. At present, web spiders mainly have two strategies: depth first search (DFS) and breadth first search (BFS).

2.2 Multi-source data web spider

In this paper, the crawling of university Internet public opinion data is under the condition that the initial URL is fixed, and the crawling is also the text content of a specific topic. Therefore, it is better to choose the breadth first search strategy.

Aiming at the multi-source of Internet public opinion in colleges and universities, this paper analyzes the data from online forums, post bars, micro-blogs, news and other data source via web spider. For the data of post bar and news, login is not required, and page data can be directly obtained; for the data of online forums and micro-blogs login is required and cookies need to be configured to simulate the process. The web spider obtains the whole web page data, so it needs to extract the specific content data in combination with HTML tags and XPath and formulate different parsing strategies for the data from different sources. The flow chart of multi-source data web spider is shown in Figure 2.

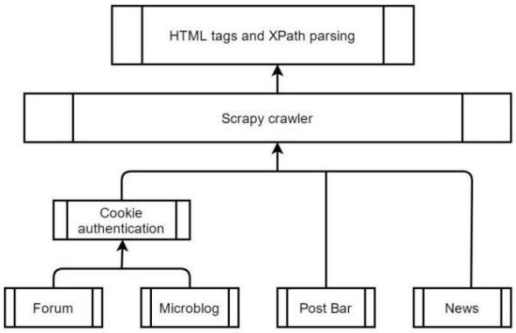

Fig. 2 Multi-source data web spider

3. Data processing and data analysis

3.1 Data extraction

The web spiders crawls the online forums, post bars, micro-blogs, news and other web page data of colleges and universities, and obtains the whole web page data containing HTML tags, including HTML header statements, JavaScript and CSS style which have nothing to do with information. But in fact, the Internet public opinion content only accounts for a part of it. The Internet public opinion content is obtained by web page parsing technology. The parsing of HTML tags is based on DOM.

Page parsing is to extract the required data from HTML pages and save it in a standard format for subsequent analysis and processing. This study adopts the regular expression and formulates information matching rules based on extracted information, carries out regular expression matching on the document, and saves the matched information. This method has high requirements for regular expressions. It is necessary to be familiar with the page content format. Many third-party libraries can help complete this task. In the specific extraction process, rules should be formulated in combination with the page content to ensure sufficient analysis data.

3.2 Data preprocessing

The public opinion text data is obtained through the web spider, but data cannot be directly recognized by the computer for analysis, so it is necessary to carry out Chinese word segmentation and remove stop words and other operations. Chinese word segmentation is an indispensable part of public opinion analysis. Unlike English word segmentation, Chinese word segmentation divides Chinese text into discrete words according to certain rules. The result of word segmentation also directly affects text processing. This paper adopts Jieba as word segmentation, an open source, portable and fully functional word segmentation tool. After word segmentation of the data from online forums, post bars, micro-blogs, and news, for some words that do not support topic detection, as well as some connectives, modal particles, auxiliary words, prepositions, and punctuation should form a stop list and these words should be filtered. This paper adopts the stop list of Harbin Institute of technology and Baidu.com.

3.3 Topic detection

Text digital vectorization must be done to the text data from the web spider. In the space vector model, we pay more attention to the similarity of texts in direction, so cosine similarity is adopted to reflect the similarity between texts. The basic steps of this method are: first, the training set and the text to be classified are represented by vectors, then the similarity between the text to be classified and the training set is calculated, then the categories of several most similar texts are calculated, and finally the categories of the text to be classified are determined.

Topic detection is to detect topics from unknown text data. Unsupervised clustering algorithm is usually used to detect topics. The clustering algorithm can merge unknown texts into several clusters, so that each cluster can basically express the same meaning, and therefore the originally unrelated texts are connected. The decision-makers can obtain the implicit knowledge of the text through clustered information.

In the traditional university network public opinion topic detection, the frequent used clustering method is K-Means clustering [2]. K-Means clustering has a simple operation and low complexity, which is suitable for processing under large data. But the disadvantage is the number of clusters needs to be set in advance, and the clustering results depend on the initialized cluster center to a certain extent. It is sensitive to noise data, which leads to local optimization. According to the characteristics of university Internet public opinion data, this paper adopts the Single-Pass clustering algorithm [3]. When using the algorithm for topic detection, each piece of data to be processed should be compared with the previously processed data to calculate the similarity between them, so as to judge the topic classification. The process of single-pass clustering algorithm is shown in Figure 3.

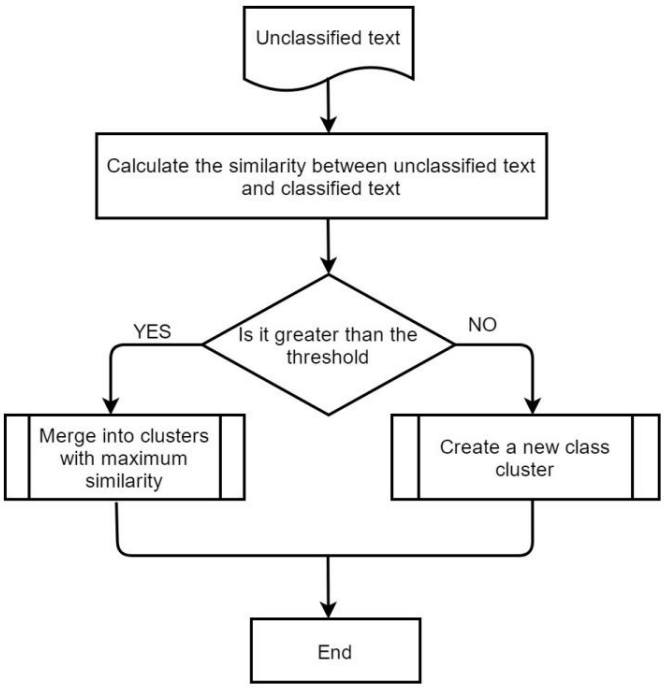

Fig. 3 Single-Pass clustering algorithm flow

3.4 Emotional analysis

Emotion analysis is an important part of Internet public opinion analysis. It is used to understand the emotional tendency of Internet users on an event. Emotion analysis can play an important role in guiding public opinion. The emotional tendency in Internet public opinion refers to the public's attitude tendency towards a social hot event in the virtual space. It can generally be reflected in two ways: one is the polarity of emotional tendency (positive and negative), and the other is the extreme value of emotional tendency (i.e. degree).

At present, there are a lot of research on emotion analysis technologies. We use Support Vector Machine algorithm. Support Vector Machine (SVM) is a classifier, which can separate samples of different classes in the sample space. The separation plane generated between the two is called the separation hyperplane. Using some labeled training samples, SVM can output an optimal separation hyperplane. The essence of SVM algorithm is to find a hyperplane to maximize a value, that is, to find the minimum distance between the hyperplane and all training samples - interval. Therefore, SVM is used to fuse emotional feature vectors to judge the emotional polarity of text. The main steps of emotion analysis are shown in Table 1

Table 1. Main steps of emotion analysis

1) Broadcast the emotion dictionary to each slave node through the master node, so that each slave node can use the same emotion dictionary.
2) Connect and convert the calculated vector and emotion score in the data processing module into LabelPoint type.
3) The transformed data are randomly divided, of which 70% are used for training and 30% are used for testing. The support vector machine classifier with maximum iterative step size of 1000 is used for data fitting.
4) Finally, the training model is used to predict and classify the test data, and the emotion analysis results are output.

4. Design and implementation of public opinion analysis system in Colleges and Universities

4.1 System design

The Internet public opinion analysis system of colleges and universities mainly aims at the data collection, pre-processing, analysis and processing, visual display. The main sources of data are school forums, post bars, micro-blogs, news, etc. Through the analysis of these public opinion data, it can help the school effectively supervise the guidance of campus public opinion and respond to unexpected events in time, to form a good campus environment.

According to the actual situation of Internet public opinion analysis in colleges and universities, this study is designed from the aspects of data collection, data processing, data analysis, front-end display and so on. The system flow is shown in Figure 4.

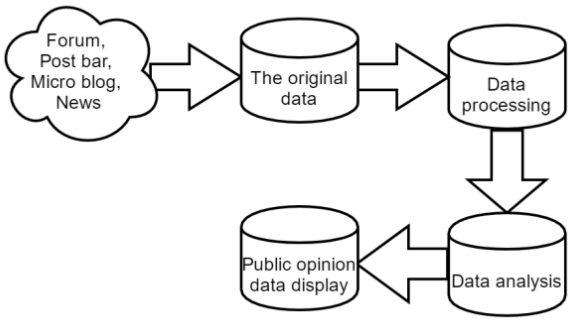

Fig. 4 Public opinion analysis system process

4.2 System implementation

Each large module is composed of different small modules. The data acquisition module includes data extraction, preprocessing, etc. the data processing module includes text word segmentation, text topic detection, emotional tendency analysis, etc. the front-end module presents the data analysis

results to users after visual processing, which enables users to collect the key information of Internet public opinion more quickly and effectively.

5. Conclusion

Internet public opinion analysis plays an important role in the cyber society. People express their views through the Internet. With the advent of the Internet big data era, the traditional public opinion analysis methods are no longer applicable, and the big data method needs to be used for public opinion analysis.

According to the actual needs of Internet public opinion analysis, mainly according to the characteristics of university public opinion analysis, this paper adopts the suitable algorithm and designs a university Internet public opinion analysis system based on web spider and modern natural language processing, which has a certain significance for university public opinion management and response.

Internet public opinion analysis is not only an important means to maintain social stability and improve the effectiveness of social management, but also a huge systematic project. Due to my capabilities, there are some deficiencies in the design of this system, which needs to be further improved and refined in the follow-up work. Finally, a more rapid, efficient and comprehensive university Internet public opinion analysis system is realized.

References

[1] Allan J, Harding S, Fisher D, et al. Taking Topic Detection From Evaluation to Practice [C] //Hawaii International Conference on System Sciences. IEEE, 2005:101a-101a.

[2] Zhang D, Li S. Topic detection based on K-means [C] // International Conference on Electronics, Communications and Control.IEEE, 2011:2983-2985.

[3] Yi X, Zhao X, Ke N, et al. An improved Single-Pass clustering algorithm internet-oriented network topic detection [C]//Fourth International Conference on Intelligent Control and Information Processing. IEEE, 2013:560-564.

[4] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer Verlag, 1995.

[5] Allan J, Harding S, Fisher D, et al. Taking Topic Detection From Evaluation to Practice[C]//Hawaii International Conference on System Sciences. IEEE, 2005:101a-101a.

[6] Andrew McCallum, Kamal Nigam. A comparison of event models for naive bayes text classification [J]. AAAI-98 workshop on learning for text categorization, 1998: 41-48.

[7] Allcott H,Gentzkow M. Social Media and Fake News in the 2016 Election [J]. Journal of economic Perspectives,2017,31(2):211-235.

[8] Sun Liwei, He Guohui, Wu Lifa. Research on web crawler technology [J]. Computer knowledge and technology, 2010, 06 (15): 4112-4115.

[9] Li Tianzhu. Design and implementation of university network public opinion analysis system [D], Chongqing University, 2021.

[10] Gong Hechen. Research on text public opinion analysis algorithm based on wechat public platform [D], Beijing Institute of printing, 2020.

[11] Ye Yongtao. Hot topic extraction and tracking based on Chinese microblog [D], Xihua University, 2017.

[12] Li Wenkun. Research on new word discovery and topic detection technology for microblog [D]. Beijing University of information technology, 2015.