Analysis and Research on the Construction of College Spirit under Big Data Technology

Ling Wang

wlcjsmxy2022@163.com

School of Finance and Trade, Chongqing Aerospace Polytechnic, Chongqing 400000, China

Abstract. In the digital information age, new media technology has influenced college students' study, life and way of thinking with its unique connotation and characteristics, thus opening up new channels and methods for the construction of college spirit. At the same time, colleges and universities should also pay attention to the monitoring and handling of online public opinion, respond to negative public opinion in time, and protect the reputation and image of the school. In view of the lack of effective monitoring of campus network public opinion in the process of building school spirit, this paper puts forward a set of real-time analysis system based on big data to improve the early warning mechanism, so as to reduce the negative impact, maintain a good school spirit and maintain a good campus atmosphere. The system is based on Hadoop framework, with Spark as the real-time analysis framework of public opinion, and combined with web crawler, text processing, data visualization and other technical means to realize the collection, preprocessing, storage, analysis, mining, visualization and other operations of campus network public opinion data, thus providing an important basis for the judgment of campus network public opinion. Practice has proved that the system has outstanding performance in distributed storage, distributed computing, text content mining, centralized display and so on, which can meet the actual needs of college spirit construction and has certain promotion significance.

Keywords: Big data; school spirit construction; network public opinion; Spark; text mining; computer application.

1 Introduction

In the new stage of innovation and development of modern education, the construction of college spirit, as the breakthrough and support point of deepening reform in colleges and universities, is gradually improving the quality of talent education in colleges and universities by virtue of its functions of encouragement, enlightenment, restraint, protection and extension, and providing the necessary environment and atmosphere for the integration and breakthrough of various special disciplines. [1] With the maturity and wide application of digital information technology, new media technology has become the main channel for contemporary college students to obtain information and communicate, which has promoted the change of college students' study, life and way of thinking, enriched the campus cultural life of colleges and universities, and made it show the characteristics of diversity, dynamics and interaction. [2] At the same time, the construction of school spirit in colleges and universities will also actively take advantage of new media technology to effectively enhance the influence and image of the school by publicizing the advantages of the campus,

establishing an online social platform, guiding the spread of positive energy, and strengthening the construction of online educational resources, so that the construction of school spirit can be more deeply rooted in the hearts of the people. However, in the actual application process, colleges and universities lack effective monitoring and handling methods for negative information and negative public opinion derived from new media technology, and the management mechanism is obviously lagging behind, which has a negative impact on teachers and students, thus weakening the guiding function of campus culture and bringing a huge impact on the construction of school spirit. In view of this, this paper believes that colleges and universities should pay attention to the monitoring and handling of campus network public opinion, take effective measures to deal with negative information and negative public opinion in time, establish a sound supervision mechanism, and avoid all kinds of risks in order to protect the image and reputation of the school and maintain a good campus atmosphere. [3] The real-time analysis system of big data can give full play to the practical advantages of web crawler, data management, text processing, data visualization and other technical means, and form a set of practical and effective comprehensive application solutions to promote the judgment of online public opinion in colleges and universities, and provide some ideas and references for the construction of school spirit in colleges and universities.

2 System construction

According to the actual requirements of the system and technical system standards, the overall design of the real-time analysis system of online public opinion in colleges and universities based on big data involves two parts: structural framework construction and functional application development. Firstly, the basic structure of the system can be divided into three parts: storage layer, middle layer and presentation layer. [4] The storage layer relies on HDFS, HBase distributed database and Mysql under Hadoop to store the obtained data and the results of real-time analysis and processing. The middle layer is the application server part of the system, which is responsible for the business logic control of the system. The presentation layer is realized by Javaweb technology, which can provide users with an interactive interface to support the display of various data information.

Secondly, the operation logic of the whole system is shown in Figure 1. In the specific development process, Hadoop data management cluster, as the application foundation of big data technology, needs to select hardware devices and software programs according to the amount of data processed by the system. This Hadoop framework needs three functional nodes, named Master1, Slave1 and Slave2 respectively. Each functional node is equipped with a server, and the hardware configuration includes two 4-core CPUs with a main frequency of 2.5GHz, 4GB of memory and 200G hard disk to meet the distributed storage requirements of data. As for the software program, Linux is selected as the bottom operating system of each server, the version is CentOS 7.5, jdk1.8.0_171 is selected as the JDK version, and Hadoop is selected as v2.6.0, and components such as Spark, Yarn, HDFS, Zookeeper, HBase, Kafka and Redis are also deployed in each node. [5] In addition, the WebCollector-Hadoop version is selected for the WebCollector crawler framework, and it is directly deployed on the Hadoop framework to realize the distributed collection of online public opinion data information. The text processing module completes Jieba word segmentation and TF-IDF feature operation under the Spark framework.
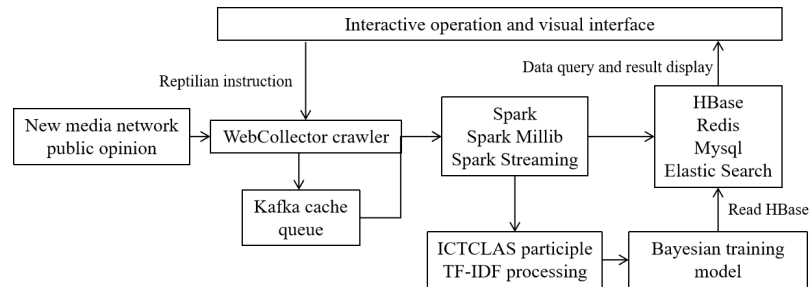
Fig. 1. Flow chart of system operation

Finally, Java is selected as the design and development environment for the interactive operation and visual interface of the system client, Apache 2.4 is selected as the Web server, IntelliJ IDEA is selected as the Java integrated development tool, and MySQL is selected as the database. This system uses Maven 3 to complete the project structure management, divides the whole project into several engineering modules, and completes the integration and encapsulation of each functional module based on Spring architecture, and opens various application API interfaces to users. [6] After the installation and configuration of the above software systems one by one, the final Web application is published on the server side, and the corresponding network and IP address are configured, which can support users to log in and use.

3 Detailed function implementation

3.1 Public opinion management

Under this function module, users can initiate online campus public opinion management operations for different channels and types, that is, select or add management items through the public opinion list in the interface. In the process of adding, users need to input keywords or key sentences themselves to determine the data crawling content of the system, and then determine the running direction of the system.

The public opinion items input by users will be recorded in Mysql, and the crawling address of data will be automatically generated after being queried and obtained by the task scheduling module, and will be written in Redis according to a specific format for WebCollector to read, thus triggering the crawler task. The Handler information mechanism of WebCollector crawler is simple and practical, and the content crawling function can be realized only by declaring customization and passing it to the crawler. The following is the implementation code. [7] After the WebCollector successfully crawls the data, it will be written into Kafka, paving the way for the subsequent public opinion data information processing.

```
Handler gene_handler = new Handler() {  @Override

  public void handleMessage(Message msg) {

    Page page = (Page) msg.obj;

    System.out.println(page.html);    }
```

3.2 Real-time monitoring

Real-time monitoring is the core function application of the system, and it is also the key index of campus network public opinion analysis and processing. The system will use Spark Streaming framework under Spark distributed computing engine to get crawled data information from Kafka message queue, and complete Chinese word segmentation, stop words filtering, result output and other operations with the help of Jieba tool and TF-IDF algorithm, and finally form text keywords and key sentences with obvious characteristics, and then classify emotions by using incremental naive Bayesian classifier based on Spark MLlib.

The core Bayesian law of incremental naive Bayesian algorithm is the optimization of sample incremental learning mode based on naive Bayesian algorithm, that is, the classification results are repeatedly included in the training samples, thus improving the accuracy of subsequent classification. The formula of incremental naive Bayes algorithm is shown in formula 1, where x is the prior probability in initial training, e is the prior probability of initial data, P(E|X) is the original probability density function, s represents the new data set, and P(E|S,X) is the posterior probability. During the off-line test, 200 samples were selected for Naive Bayes algorithm training and incremental Naive Bayes algorithm training, and the comparison results are shown in Tables 1 and 2. [8] The incremental naive Bayes algorithm will select the classification results with confidence greater than 80% when expanding the sample size, and re-fit the algorithm model. In the end, the classification accuracy of the two algorithms is 0.8322 for incremental naive Bayes classification, which is better than 0.7703 for naive Bayes algorithm.

$$P\left(E_i \mid S, X\right) = \frac{P\left(S \mid E_i, X\right)P\left(E_i \mid X\right)}{P\left(S \mid X\right)} \tag{1}$$

Table 1. Classification results of 5 batches of samples

	01	02	03	04	05
Number of samples	50	50	50	50	50
Naive Bayes classification accuracy	0.7680	0.7655	0.7641	0.7759	0.7784

Table 2. Classification results of 5 batches of sample increments

	01	02	03	04	05
Number of samples	50	50	50	50	50
The cumulative number of samples	50	77	97	129	149
The algorithm model	1	2	3	4	5
Incremental naive Bayes classification accuracy	0.7678	0.8177	0.8215	0.8559	0.8984

According to the results of incremental naive Bayes classification, the positive and negative values of public opinion text content are assigned, with the positive value being 1 and the negative value being 0. [9] At the same time, it will also calculate the current proportion of positive public opinion and negative public opinion, and finally input the results to Hbase for saving.

Trend analysis. According to the monitoring results of real-time public opinion, the system will combine the data visualization plug-in D3.js to transform the dynamic evolution within

the time range into the trend of public opinion, which is convenient for users to view intuitively. Figure 2 shows the trend of public opinion at a certain time in colleges and universities.

Fig. 2. The trend of public opinion at a certain time in colleges and universities

Public opinion word cloud. After the system completes word segmentation and key feature extraction of public opinion text content, it will also enable the UpdateStateByKey component to calculate the occurrence times of keywords or key sentences in real time, and finally write the results into Redis for the front-end D3.js to call to generate a word cloud map. [10]

3.3 Public opinion early warning

The public opinion early warning function of the system will be attached to the Compute Sentiment Index program, that is, the corresponding threshold will be set according to the proportion of positive public opinion and negative public opinion. Once the proportion of negative public opinion exceeds the set threshold, the public opinion early warning will be triggered, and the system will send an email to the designated mailbox through the WebAPI provided by the SendGrid component, thus realizing the early warning function. Figure 3 shows the word cloud of public opinion hot words in a certain period of time in colleges and universities.

Fig. 3. The word cloud of public opinion hot words in a certain period of time in colleges and universities

4 Conclusions

Taking the monitoring of online public opinion in the process of building college spirit as the research object, this paper constructs a real-time public opinion analysis system by means of

big data, web crawler, data management, text processing, data visualization and other technical means. The system can quickly complete the collection, preprocessing, storage, analysis, mining, visualization and other operations of campus network public opinion data, thus providing an important basis for the judgment of campus network public opinion and helping the construction of school spirit in the new media era of colleges and universities. In the follow-up research, the system will further improve the management efficiency of public opinion data information, strengthen the application of various data analysis and mining models, realize the correct guidance of campus network public opinion, and make a beneficial attempt for the comprehensive construction of college spirit in the new era.

References

[1] Wang Jiayi et al. The Soft Environment Construction of Academic Atmosphere Construction in Colleges and Universities in the New Period from the Perspective of Students' Ideological and Political Work[J]. Defence Industry Conversion in China.12 (2022)

[2] Liudmyla Mialkovska et al. New Media as Modern Communication Technologies: The Digital Dimension.[J].Khazar Journal of Humanities and Social Sciences.01 (2023)

[3] Li Dongke et al. Application of Big Data Technology in Campus Public Opinion Management[J]. Electronic Techniques.05 (2023)

[4] Jiao Xiangyu, Huang Kanghui, et al. Hadoop+JavaWeb Big Data Analysis Visualization System[J]. Management & Technology of SME.02 (2020)

[5] Archana Uriti et al. Understand the Working of Sqoop and Hive in Hadoop[J].Applied and Computational Engineering.06 (2023)

[6] Shi Feng. Development and Application of JavaWeb Based on MVC Pattern[J]. Electronic Techniques.05 (2021)

[7] Shan Wenyuan. Research and Design of Distributed Topic Web Crawler[D]. University of Electronic Science and Technology of China.03 (2020)

[8] Du Kejing. Research and Application of Three-Branch Incremental Naive Bayesian Algorithm[D]. North Minzu University.06 (2022)

[9] Wang Shihang, Tang Yanjun. Text Analysis of Internet Public Opinion Based on Emotion Classification and Topic Mining[J]. Network Security Technology & Application.07 (2023)

[10] Lin Suhong, Du Yuhong. Method of Text Keyword Extraction and Visualization Based on Comment Data[J]. China Computer & Communication.12 (2022)