Research on the Computer Informatization in Multimedia Public Opinion Monitoring

Chunshu Xu1, * and Zhongyuan Xu2
1Heilongjiang Bayi Agricultural University, Daqing 163319, China
2Heilongjiang Nongken Industrial School, Hegang 154100, China

*Corresponding author email: chunshuxu@byau.edu.cn

Abstract. This article designs and implements a Hadoop-based network public opinion monitoring system. The system uses HDFS as the underlying storage system, and builds a distributed database based on HBase on it for unified storage and management of public opinion information. The paper will introduce a Hadoop-based microblog public opinion monitoring system that can mine and analyse large-scale collected data, realize the discovery and tracking of public opinion hot topics, and the social network analysis of Weibo. The analysis results are visualized and presented for the party. Government agencies, large enterprises and other units and organizations discover sensitive information in a timely manner, grasp public opinion hotspots, grasp public opinion trends, and provide automated, systematic and scientific information support in response to public opinion crises.

Keywords: Computer, informatization, multimedia, crowd public opinion, public opinion monitoring, Weibo public opinion.

1. Introduction

In recent years, the influence of Internet public opinion on the occurrence, development and changes of social events has been increasing day by day, and the public’s supervision of public issues and social management has played an increasing role. At the same time, if online public opinion emergencies are not handled in a timely manner, improperly handled, or poorly managed, netizens use Weibo, WeChat, forums and other network tools to gather strength and create public opinion pressure, publish false information, and vent their personal emotions. It induces the negative and excessive emotions of the general public, thereby triggering excessive and even illegal behaviours of some people, posing a threat to social stability. How to integrate, analyse, and dig out hot topics on the basis of big data and large collections of online public opinion, and to track hot topics, adopt corresponding guidance processing strategies at different stages to effectively control online public opinion events, has attracted widespread attention [1]. At present, these mainstream hot spot tracking algorithms are all based on small networks. When processing massive web pages, the content is complicated and there are too many interference factors, and it is difficult to accurately mine the results of hot centre topics. Hadoop is a typical big data network based on open-source cloud computing platform. Therefore, the existing hot tracking modelling method is improved, and the research on the tracking method of network public opinion hot topic based on Hadoop can greatly improve the accuracy of mining hot topics and efficiency.
2. System Architecture
The design of the network public opinion monitoring system is a systematic project. In the system design, in order to ensure the effectiveness of the design, it is first necessary to carry out a detailed demand analysis of the system [and determine the functional architecture of the system. In terms of non-functional requirements, the Internet generates massive amounts of data every day. Take Weibo as an example [2]. Weibo has more than 400 million active users, more than 1,000 Weibo are generated every second, and the daily new data volume is more than 10TB. It can be seen that the difficulty in the design of the public opinion system lies in the design of the storage platform. The monitoring system needs a storage platform with strong availability and high fault tolerance. When a server in the storage platform is down, the stored data can still ensure its integrity; secondly, the information obtained based on web crawlers is not in the data format. Unified, the storage platform must ensure the consistency of the storage format. In terms of functional requirements, starting from the goals that the system needs to achieve, it is disassembled level by level into the functional architecture diagram shown in Figure 1.

![Diagram of Hadoop-based network public opinion analysis system structure](image)

**Figure 1.** Hadoop-based network public opinion analysis system structure diagram

2.1. Data acquisition module
Refers to the collection of webpage data based on web crawler technology or the data collection API of social networking sites. This article mainly relies on web crawler technology. The specific process of crawler technology is to first obtain the URL of the initial webpage and store it in the webpage queue, obtain the source code of the webpage, extract the effective information on the webpage, and store it in the database.

2.2. Data pre-processing module
After collecting the data on the web page, the standard Jason format web page data information is obtained. In order to ensure the accuracy of the results, it is necessary to perform data pre-processing through filtering, deduplication and other operations, and establish a text model. The specific process includes: 1) Remove useless information in the text. 2) Word segmentation. A dictionary is established and word segmentation is performed on the text. In the process, new words found in the text are added to the dictionary using cascading hidden Markov chains, and the dictionary is constantly updated to improve the accuracy of word segmentation. 3) Remove stop words. Due to the particularity of Chinese, there are many conjunctions and modal particles. These words need to be removed before text analysis, otherwise it will increase the burden of text analysis algorithms and reduce the accuracy of processing.
4) Build a text model. Use Dirichlet Model (LDA) to associate text information with topics to build a text model.

2.3. Topic Discovery Module

Use clustering algorithm K-means+ as the text distance algorithm of the system. The algorithm can process large-scale data, and has strong scalability and robustness to noisy data. Due to the large amount of data, it needs to be parallelized based on MapReduce. The process of parallelization includes 3 steps: 1) Map process is to convert the output into multiple <key, value> key-value pairs; 2) Combine process is to average the values with the same key on a single node to reduce the network the amount of data transmission on the server and reduce the storage burden of the server; 3) The Reduce process is to average the value of each node with the same key. After the text is clustered, hot topics can be found according to the attributes of the text, such as forwarding, comments, and likes. The article comprehensively considers the number of reports, the source of the report, the speed of the report and other factors, and uses the formula (1) to calculate the hot topics:

$$focus = \frac{r_n \times N}{R_N} + \frac{r_d \times N}{\sum_{i=0}^{N} (r_d^i)} + g \log \left( h_n + s c_n \right)$$  \hspace{1cm} (1)$$

Among them, $r_n$ is the number of relevant reports during this period; $R_N$ is the total number of all reports during this period; $N$ is the total number of topics during this period; $r_d$ is the total number of reporting days; $\sum_{i=0}^{N} (r_d^i)$ is the total number of reporting days for all topics; $h_n$ is The number of topic clicks; $c_n$ is the number of topic comments. The specific implementation process is shown in Figure 2.

![Figure 2. Hotspot calculation process](image)

The map function has two parts. The first part is to calculate the hot spot (calculate the above 1, 4, 6, 7 items, 2, 3, 5 cannot be calculated because other types of data are needed). Pass the calculated items 1, 4, 6, and 7 to the reduce function to calculate the remaining 2, 3, and 5 items, and finally calculate the hot spots.
2.4. Public Opinion Analysis Module
It is to analyse the spread path of public opinion, the tendency of the topic, and the social attributes of the topic. Among them, the analysis of the propagation path relies on the forwarding network established according to the time series during the crawling process [3]. The tendency analysis of a topic relies on the analysis of the sentiment words of the text information after the text clustering, combined with its weight to determine; the social attributes of the topic mainly rely on the nodes obtained by the topic, so as to determine the person attributes related to the topic.

3. Experimental test
Hadoop-based microblog information data uses API-based (application programming interface) and a structure tree webpage text extraction and parsing solution from massive microblog data. The obtained microblog data is stored in a Hadoop-based text database. Realize parallel and efficient pre-processing, in-depth analysis and real-time warning of public opinion [4]. The microblog public opinion analysis and early warning system framework based on Hadoop is shown in Figure 3. The key technologies of its module realization are as follows.

![Figure 3. Hadoop-based microblog public opinion analysis and early warning system framework](image)

3.1. Experimental environment
In terms of hardware, the experiment consists of 4 ordinary PCs of the same structure connected through a switch to build a small Hadoop cluster. The 14.04 Ubuntu operating system is installed on the 4 PCs respectively, and Hadoop 2.2.0 is deployed on the Ubuntu system to complete the Hadoop cluster. Set up, use one of the PCs as the master node and name it Host Master to start the Job Tracker and Name Node processes. The remaining three machines are named slave1, slave2, and slave3 as slave nodes to start the Task Tracker and Data Node processes [5]. The IP addresses of the 4 PCs are 172.30.78.1-172.30.78.4 respectively. The specific parameters of the server and switch are shown in Table 1 and Table 2.
Table 1. Server parameters

| Name              | Parameter                      |
|-------------------|--------------------------------|
| Model             | Dell PowerEdgeR730             |
| CPU               | XeonE5-2609v3                  |
| RAM               | 64GBDDR31333MHz                |
| Hard drive capacity | 2TB                            |
| Operating system  | CentOS6.6x64                   |

Table 2. Switch parameters

| Name             | Parameter                          |
|------------------|------------------------------------|
| Model            | Huawei S5700-24TP-SI(AC)           |
| Product type     | Gigabit Ethernet Switch            |
| Application level| 3 layers                           |
| Port             | 24 10/100/1000Base-T ports, 4      |

When configuring the server, the IP configuration of the 13 servers in the architecture is between 192.168.1.1~13. Configure one of the servers as Master mode, and configure the remaining 12 servers as Slaver1~Slaver12. Then proceed with SSH configuration: first produce a public key and a private key for each server node, then start SSHkey to perform a stand-alone loopback test, and then save the public key of the Master node into the public key of Slaver, so as to realize the password-free login of the Master node.

3.2. Test

3.2.1. Discovery and visualization of hot topics. The user reads the clustering centre and its contained sub-items in the analysis library through the interactive module, and then uses the force-oriented algorithm and JavaScript to visualize, as shown in Figure 4: The centre point is the topic of the day, and the small dots around it represent participation the blogger on this topic.

![Figure 4. Visualization of hot topics](image-url)
3.2.2. Social network analysis. The three most important characteristics of complex networks are small-world characteristics, scale-free characteristics, and high clustering coefficients. This system analyses the scale-free and small-world characteristics of Weibo networks through node degree distribution and clustering coefficients. After performing MapReduce calculation on the attention and fan relationship in the information, the result is shown in Figure 5. The upper figure is the node out-degree distribution graph, and the lower figure is the node in-degree distribution graph [6]. The upper graph in Figure 5 shows that the out-degree is There will be a spike around 160, and the statistics show that there are 92 users with a degree of between 150 and 160. From the figure below, it can be seen that there are more than 200 followers with more than 1,000 followers, of which 62 people posted Weibo at 100 Above, so these users are active users. They not only have a large number of friends, but also have many followers, and often post Weymouth’s users are the most active groups in the network.

![Figure 5. Visualization of social network analysis](image)

4. Conclusion
Aiming at the rapid development of Weibo, a social network, this article proposes a Hadoop-based Weibo public opinion monitoring system model, and studies the application of Hadoop distributed storage and MapReduce parallel computing models in a massive Weibo public opinion monitoring and analysis under a big data environment. The workflow and implementation of the model components are designed in detail. Through experimental simulation, the Hadoop-based microblog public opinion monitoring system can effectively monitor and analyse large-scale microblog data, try other clustering algorithms, perform comparative analysis, and complete the microblog public opinion monitoring system based on Hadoop platform to obtain hot topics accuracy.

Acknowledgments
The source of the results of this paper is: "Heilongjiang Agricultural University Support Program for San Heng San Zong", the project number is: RRCQC202004, and the project name is: Research on the Spread Effect of "Longjiang Spirit" in the All-Media Age.

References
[1] Chen, J., & Yang, J. Maximizing coverage quality with budget constrained in mobile crowd-
sensing network for environmental monitoring applications. Sensors, 19 (10) (2019) 2399 - 2409.

[2] Taha, A. E. M., & Abu Ali, N. Monitoring a crowd's affective state: status quo and future outlook. IEEE Communications Magazine, 57 (4) (2019) 26 - 32.

[3] Hao, Wang, Dong, Zhao, Huadong, & Ma. Informative image selection for crowdsourcing-based mobile location recognition. Multimedia Systems, 25 (5) (2019) 513 – 523.

[4] Li, L., Wen, G., Wang, Z., & Yang, Y. Efficient and secure image communication system based on compressed sensing for iot monitoring applications. IEEE Transactions on Multimedia, 22 (1) (2020) 82 - 95.

[5] Shao, Y., Li, W., Chu, H., Chang, Z., & Zhan, H. A multitask cascading cnn with multiscale infrared optical flow feature fusion-based abnormal crowd behavior monitoring uav. Sensors, 20 (19) (2020) 5550 - 5559.

[6] Gu, Y., Shen, H., Bai, G., Wang, T., & Liu, X. Qoi-aware incentive for multimedia crowdsensing enabled learning system. Multimedia Systems, 26 (1) (2020) 3 - 16.