Network Public Opinion Monitoring Based on MapReduce

Yunhong Zhao and Suruo Li
Jingchu University of Technology, Jingmen, HuBei, China

*Corresponding author e-mail: zyhong@jcut.edu.cn

Abstract. Public opinion information on the network is abundantly decentralized in Internet, these false ones, easily misleading, not legally binding, it is difficult to deal with. Quickly master information is a prerequisite to deal with the network of public opinion, clustering algorithm in data mining plays an important role in in the collection of information statistics. Disadvantages of classical clustering methods are high resources consumption, low efficiency, high time and space complexity, and difficult to deal with large-scale data processing in massive network. To the network public opinion text as the research object, in-depth study of summarizing the network public opinion monitoring technology based on distributed MapReduce, including parallel technology, improved clustering algorithm, relational database and distributed database construction, which is in order to improve the efficiency of information processing, to reduce the pin.

Keywords: MapReduce, Network Public Opinion, Text Clustering, Birch Algorithm

1. Introduction

With the rapid development of social networking platforms and mobile Internet, new media such as forum, community, WeChat, micro-blog become the main channel for people to obtain and exchange of information. As of December 2020, the number of Internet users in China reached 989 million, and the Internet penetration rate was 70.4% [1]. Internet has become the distribution center of Ideological and cultural information and the amplifier of social public opinion. Everyone is from the media; Internet users are willing to publish information to the network. Events that do not discriminate between the true and false, disseminate information, follow blindly, listen to the Internet rumors, and add fuel to the flames, because strong negative impact events are extremely impressive. Public opinion public opinion to deal with it properly, a direct test of the ability of enterprises, government departments and public, thus affecting the business, establish credibility and brand image of the department. Clustering algorithm is the key processing of massive data and important technology, its efficiency and parallel ability depends on the level of monitoring of network public opinion, and thus leads to serious consequences [2].

Modern network information is changing rapidly and increasing in geometric order. The data storage and calculation are changed from local to cloud, and the storage form is changed from centralized to distribute. The typical traditional clustering algorithms are limited in the face of distributed large-scale network data, especially in the aspects of resource cost, efficiency and complexity [3]. It is particularly necessary to design a clustering algorithm and technology suitable for
parallel processing of large data, which is of great significance for the rapid clustering analysis and decision execution of online public opinion.

2. Research Status and Analysis
Because most of the network public opinion information is in the form of text, so only text clustering is researched and analysis summary.

2.1 Research and Application of Text Clustering Mining Abroad
Western and European developed countries have carried out extensive research on text clustering mining, and the earliest literature began in 1957. The University of Michigan Radev et al first proposed the concept of mass, recommendation system design of MEAD, with automatic extraction and identification of useful information, the working mode is automatic clustering daily news network content of the formation and provided to the user; MultiGen automatic text summarization system based on M.C. Known et al. Development of the Columbia University, in addition to automatic clustering document summary, function in the removal of redundant information is more powerful. The automatic document system developed by Ming Hui Wang and Hedihekotanaka of University of Tokyo has been clustered according to references; because the application fields of Chinese information processing are relatively narrow [4]. Clustering is used in business, the development of Intelligent Miner for Text IBM, with text mining functions with high efficiency, can realize the feature extraction, text analysis, Web document classification and retrieval, which can support Boolean logic conditions (full text search or index) also supports natural language search, effectively improve the quality of information query. The United States Megaputer company developed TestAnalyst neural network based on Texttractor text mining search system, with a strong natural language, structured text search, thesaurus creation and other functions; the German Xtramind company with the French company Thales joint research and development of XM-Mindset in a text clustering search engine (XM-Clustering), (XM-Finder), index (XM-Index Navigator) (XM-Duplicate Recognition), redundancy identification, language identification (XM-Language Identification) (XM-Info tract), information extraction, automatic text summarization (XM-Summary) and other functional modules, has the characteristics of dynamic portability and high efficiency, precision and stability. In addition, the well-known Java based open source clustering search engine Carrot2 can automatically classify the natural search results into the corresponding semantic categories, providing hierarchical, synonymous, tag filtering functions [5].

2.2 Research and Application of Text Clustering Mining
Compared with foreign countries, the research and application of text clustering in China started late, and mainly stay in research institutes and universities. The earliest literature began in 1975. Because of the continuity of Chinese word formation and the complexity of stop words (such as Chinese characters are not like English words, there is space division, more function words, etc.), Chinese Text Clustering Mining encounters more technical problems and opportunities.

In theoretical research, Professor Wu Lide of Fudan University put forward a new indexing method for midpoint data in high dimensional space. Professor Zhou Zhihua studied the rules extraction of neural networks, and Professor Huang Changning studied the word segmentation in Chinese information processing. In addition, the four layers architecture of Internet public opinion information mining is put forward by Du Yi of Harbin Institute of Technology [6]. Professor Li Xiaoming of Peking University has studied the grid monitoring system based on distributed hash table.

In the field of applied research, many university research institutions and companies have developed and designed a Chinese text mining clustering system. A typical representative of Chinese Academy of Sciences Institute of computing technology, digital room research and development of Chinese lexical analysis system ICTCLAS mining, such as statistical analysis, feature extraction, clustering, classification and prediction of good performance in data; Harbin Institute of Technology Information Retrieval Laboratory designed Chinese multi document automatic text summarization
system prediction and Chinese text clustering system [7]. TRS data management system research and development Beijing TRS Information Technology Limited (TRS Hybase Management System) is compatible with the Hadoop standard, PB level support for massive data management and processing technology of Chinese text clustering. The disadvantage is that the search text clustering cannot provide users with navigation and other functions [8]. The research and application of text clustering mining in China has broad space for development, and it has a long way to go.

3. Network Public Opinion Monitoring Based on MapReduce

To cluster massive public opinion texts, that is, the improved Birch algorithm is used to compress the data set efficiently, and the clustering processing of the data set is carried out quickly. To rapid clustering of public opinion text in the context of high real-time requirements that is to build a hybrid database and use distributed processing technology to handle public opinion big data set [9].

3.1 Constructs the Programming Model

To the network public opinion text as the research object, in-depth study of summarizing the network public opinion monitoring technology based on distributed MapReduce, including parallel technology, improved clustering algorithm, relational database and distributed database construction, in order to improve the efficiency of information processing, to reduce the pin [10]. The project focuses on the distributed MapReduce technology, constructs the programming model, and studies the text clustering and parallelization technology in the network public opinion [11]. BRICH presents a hierarchical clustering algorithm based on distance improved, combined with distributed technology MapReduce, clustering algorithm in network public opinion in text mining and parallel processing, and the classic public opinion big data were compared to the clustering of text mining, text clustering to improve the parallel performance of public opinion, improve the effectiveness and real-time, through experiments verify the conclusion is more suitable for clustering processing of massive network public opinion information.

![Figure 1. Network public opinion monitoring model](image)

Through the construction of network public opinion monitoring model, such as Figure 1 to achieve network data collection.

3.2 Key Technologies and Methods

The related technology of each stage of public opinion summary text clustering, text preprocessing (including Chinese segmentation, filtering of stop words (Stop, Word)) feature extraction (TFIDF
algorithm), representation model (Boolean model, vector space model, probability model, space model), similarity calculation (Jaccard similarity method). Then, we summarize the typical clustering algorithms (including hierarchy, partition, density, mesh, model, constraint, fuzzy, granularity, quantum, kernel, spectrum), and finally study the evaluation criteria.

Analysis of the advantages and disadvantages of Birch algorithm, the improved clustering feature tree CF-Tree (Clustering Feature Tree) in the construction process; improve the design method of Birch algorithm: threshold setting, outlier mining and removal.

In a distributed parallel computing framework Hadoop/MapReduce, using a text clustering algorithm Birch is improved by JAVA program, compare the parallel performance of Birch text clustering algorithm and the effectiveness of parallel text clustering is more suitable for the treatment of massive network public opinion information verification conclusion.

3.3 Simulation Experiment
The experimental data source is based on the Chinese text of Sohu network Sougu laboratory, the natural language classification corpus Sougu-T (10 kinds of cars, etc.). The platform is free to use only for registration, and its corpus is mainly used in the research of standard Chinese clustering and Natural Language Processing.

In this experiment, 6 sets of Hadoop cluster system is built on the VMWare10/Ubuntu12.02 platform using Eclipse3.7 Java development environment; integrated environment; cluster environment is: 1 Master virtual machine is responsible for task scheduling, 5 worker virtual machine implementation of map and reduce. cluster node with planning and design of Table 2 cluster node planning.

| Item             | configuration                        |
|------------------|--------------------------------------|
| CPU              | Intel(R)Core(TM)i5-6402P@2.80GHz 2.80GHz |
| Memory           | 1GB                                  |
| Hard disk        | 20GB                                 |
| Virtual machine OS | Ubuntu12.02                          |
| Virtual machine system | VMWare10                        |

Table 1. Virtual machine server configuration

| Development tools and platform | Eclipse3.7,JDK1.7,Hadoop |

Table 2. Node planning of cluster environment

| host name | IP address | HDFS node | MapReduce node |
|-----------|------------|-----------|----------------|
| Master    | 218.193.154.16 | NameNode,DataNode | JobTracker |
| Slave1    | 218.193.154.11 | DataNode    | TaskTracker   |
| Slave2    | 218.193.154.12 | DataNode    | TaskTracker   |
| Slave3    | 218.193.154.13 | DataNode    | TaskTracker   |
| Slave4    | 218.193.154.14 | DataNode    | TaskTracker   |
| Slave5    | 218.193.154.15 | DataNode    | TaskTracker   |

Parallel Birch clustering based on corpus text sets is validated in a clustered environment.

4. Conclusion
The network public opinion is not the direct data in the network, but the result extracted from the sea data through related technology. Due to the timeliness reasons, the traditional network public opinion monitoring based on the accuracy priority cannot meet the demand, and the big data processing engine MapReduce technology can complete the timely processing of massive data.

The relational database (RDBMS) with standardized, structured query and data integrity and performance advantages, but in fault tolerance, flexibility and scalability exposed the shortcomings. The distributed database (HDFS) not only has good parallel processing ability, but also has high scalability and fault tolerance. In the network public opinion monitoring, to build a hybrid database, the upper RDBM is responsible for querying and processing, the upper HDFS task decomposition and
scheduling, give full play to their advantages, improve the processing performance of the network public opinion text mining.

Acknowledgements
This research was supported by Big Data Visualization Scientific Research Team of Jingchu University of Technology (Grant No.T202015).

References
[1] China Internet Network Information Center (CNNIC) 47th statistical report on Internet development in China. http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm(in Chinese)
[2] Chitrakant Banchhor, N. Srinivasu. Holoentropy based Correlative Naive Bayes classifier and MapReduce model for classifying the big data. Evolutionary Intelligence 6,19-27(2019)
[3] Nur Rokhman, Amelia Nursanti. The MapReduce Model on Cascading Platform for Frequent Itemset Mining. Indonesian Journal of Computing and Cybernetics Systems 22, 12-19(2018).
[4] He Yulin, Huang Zhezhi. Research progress of clustering algorithm for large scale data sets. Journal of Shenzhen University (Science and Engineering Edition), 36 (01),4-17(2019) (in Chinese)
[5] Yi Liu, Guiwu Wei,Haobin Liu,Lei Xu. Group decision making for internet public opinion emergency based upon linguistic intuitionistic fuzzy information. International Journal of Machine Learning and Cybernetics, 2021
[6] Tang Tao. Research on method of data analysis based on network public opinion. Modern intelligence 34, 3-11 (2014). (in Chinese)
[7] Luo Hao. On the application of network public opinion monitoring system in social public security. China Science and technology information 17, 98-102 (2011). (in Chinese)
[8] Xia Yixue, Yuan Ye, Zhang Wencai, et al. Research on abnormal data monitoring and application of network public opinion for big data. Modern intelligence, 038 (006): 80-85(2018). (in Chinese)
[9] Zhao Liang, Yu Haiyang, Li Jiao, Wei. Meteorological disaster information collection and processing in the Internet Era. Science and technology information 15, 4-5 (2017). (in Chinese)
[10] Wang Ning, Zhang Ming, Li Xiaoming. Grid monitoring system based on distributed hash table. Journal of Huazhong University of Science and Technology (Natural Science Edition) 1, 112-115 (2006). (in Chinese)