Network Big Data Mining Algorithm Based on Association Rules of Computer Technology

Yufang Jiang*
Guangxi Modern Polytechnic College, Hechi, Guangxi, China, 547000

*E-mail: jiangyufang_1118@163.com

Abstract. With the development of cloud computing, it has provided a platform for big data presentation and sharing. These data often contain artificially added uncertain factors to prevent privacy leaks. How to perform mining of these big data is a problem to be solved urgently in big data sharing. In the big data for sharing, it is implemented through the generalization process of precise data, with the feature of uniform distribution, which is not conducive to exact query, but it can provide convenience for the mining of association rules. Firstly, based on the possible intersection or inclusion relationship, hierarchical clustering is performed on the generalized values. To save the essential information related to big dataset mining, we proposed an algorithm for building an uncertain frequent pattern tree. An improved algorithm of network big data mining based on association rules of computer technology is proposed in the big data environment, and its application in landslide monitoring and early warning is discussed.

Keywords: Big Data, Association Rules, Computer Technology, Improvement

1. Introduction
Cloud computing provides a platform for presenting and sharing big data. Through the analysis and mining of big data, we can better discover the value of data and improve the accuracy of prediction. For the purpose of privacy protection[1-2], big data often contain artificially added uncertain factors. The existing data mining technology is inefficient or infeasible when applied to such big data. How to perform big data mining effectively is a problem to be solved urgently in the big data sharing at present. In the big data used for sharing, big data has the feature of uniform distribution through the generalization of precise data[3-4], which is not conducive to exact query, but provides convenience for the mining of association rules. Association rule mining is the most basic analysis method in the field of data mining. The purpose of this algorithm is to find the valuable association relationship between item sets or attributes from a large number of data. With the powerful storage and computing ability of cloud computing, this paper proposes a new algorithm for mining association rules among attributes of big data that meet the structural and uniform distribution characteristics after generalization. This algorithm sets different minimum support and minimum confidence for different generalized values[5-6]. To improve the efficiency of data mining, we first divide big data into smaller data sets according to the generalized values and then perform mining tasks on different computing nodes in parallel.
In this paper, the cloud computing distributed platform Hadoop is used to provide storage and computing processing capabilities for massive dynamic data. Given the shortcomings of traditional association rule data mining algorithms, the algorithm is optimized by using the parallelization based on computer technology, and the Hadoop high-density server cluster is used to quickly find the hidden rules that cause landslides for processing massive monitoring data. To achieve the task of extracting valuable information, to improve the efficiency of the association rules of network big data mining algorithm network big data mining in dealing with massive data, to provide a powerful prediction criterion for preventing natural disasters.

2. Network big data mining algorithm based on association rules of computer technology
With the application of Internet of things, mobile Internet, 5g and other technologies, the data generated is massive, which makes the traditional network big data mining algorithm of association rules cannot meet the needs of data processing, especially for the timely and effective processing of meteorological disaster monitoring data, the efficiency of data processing is particularly important. Next, the parallel algorithm based on computer technology is analyzed.

2.1. Improved algorithm of network big data mining based on association rules
In the data mining prototype system, we mainly use the improved association rule algorithm of network big data mining, the computation method of network big data mining. In this paper, we use the channel of iterative retrieval layer by layer, first scan the data set, discuss the process of \((K + 1)\) item set through the k-item set, and then generate frequent 1-item set L1. With the help of L1, we expand the retrieval of frequent itemset L2, through iterative way until the frequent itemset equals to the empty set. After several times of cyclic retrieval, the optimization algorithm is extracted:

1. Firstly, perform JOIN, and ensure that L.K-1 has candidate set parameter CK, and carry out connectivity operation;
2. Based on the related properties of network big data mining, the support statistics, pruning, and other operations are implemented. As a result, frequent set parameter LK is generated in CK. Because it is necessary to scan the database repeatedly to retrieve the associated frequent item set, the memory consumption is very large, resulting in slow computing speed and low efficiency. Therefore, based on the Hadoop cloud computing platform, the network big data mining algorithm is parallelized in this platform to improve resource utilization and computing efficiency. The parallel process of network big data mining algorithm based on association rules of computer technology is shown in Figure 1.

Its working principle is as follows: after the database is scanned first, map and reduce operations are carried out in different data nodes, and some frequent item sets are obtained by calculating the data of different nodes for many times. Then the main node Master is called to gather and confirm the actual global support and frequent item sets, thus saving data processing time, improving the performance of the server and improving the efficiency of data mining. The computing method of network big data mining based on computer technology is mainly concentrated in the algorithm layer of the system.

2.2. Main concept of improved mining method based on association rules
This method is based on the “parallel and distributed association rule mining” algorithm researched by park, Chen, Agrawal, etc. In the framework of network big data mining, in the process of generating frequent item sets by scanning the database many times, the association is calculated

3. Network big data mining algorithm analysis and optimization verification

3.1. Analysis of improved network big data mining_ML algorithm
The distributed file system (HDFS) is used to store and manage massive data, which is one of the core technologies of Hadoop. When the data is loaded into the HDFS system, the master node controls the data access of each node. The master node will divide the metadata into several data segments, and
then distribute the segmented data to each data node for storage. Finally, the data node is responsible for calling the map() function for computation according to the different needs and support of users. In computer-based technology, data node obtains candidate item set with certain support, counts candidate item set, uploads the statistical results to the master node, summarizes the same candidate item set to the same reduce node, and then judges which candidate item set meets the requirements according to the minimum support preset by the user, and finds out the frequent item set.

Given the sample data set, \( D = \{x_1, x_2, \ldots, x_n\}, x_i \in R^d, i = 1, \ldots, n \). Assuming that the first sample L is labeled \( \epsilon = (x_1, x_2, \ldots, x_i) \), the corresponding label is \( \eta = \{y_1, y_2, \ldots, y_j\} \).

The significance of each feature dimension in the classification problem is different. To some extent, the computer cybersecurity can overcome the shortcoming that the computer cybersecurity data treat each feature dimension equally. Its definition is as follows:

Sample \( x_i \) and \( x_j \) in the computer cybersecurity data are defined as follows

\[
d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)},
\]

(1)

Where \( x_i \in R^d, A \in R^{d \times d} \) is symmetric semi positive definite matrix.

According to the properties of semi positive definite matrix, the above equation can be written as follows:

\[
d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)}
\]

\[
= \sqrt{(x_i - x_j)^T L^T L (x_i - x_j)}
\]

\[
= \sqrt{(Lx_i - Lx_j)^T (Lx_i - Lx_j)}.
\]

3.2. Verification of optimized algorithm network big data mining _MR

During data mining, if the volume of data increases to a certain extent, the traditional network big data mining algorithm of association rules cannot support the normal operation, so it is necessary to seek other technologies and platforms to solve big data analysis. To deal with big data sets, we can build a system based on computer technology as support.

1) Construction of the system operation environment based on computer technology

This verification is carried out in the cloud computing simulation laboratory of a college. There are 16 tr227 servers in the training room. The prototype system uses 5 nodes installed with Linux operating system to build a Hadoop cluster environment, one as the master node and the other four as data node slave nodes.

2) Comparison of experimental data and performance analysis

In the running environment of Hadoop, the parallel network big data mining_ MR algorithm is packaged into jar file, and the synthetic experiment data set of natural disaster monitoring data is uploaded to HDFS. Finally, data comparison experiments are carried out in THE single-machine mode and map nodes 2 and 4, respectively, to calculate the corresponding acceleration ratio.

The data used in the experiment is stored in the form of files, and the support degree is 0.7. In this experiment, five nodes in the cluster are used, including four data nodes and one data node for the data test. The test results are shown in Table 1 below:
Table 1. Data of experimental results

| The volume of data (Unit: ten thousand) | Stand-alone (Unit: seconds) | 2 nodes (Unit: seconds) | 4 nodes (Unit: seconds) |
|----------------------------------------|-----------------------------|-------------------------|-------------------------|
| 0.5                                    | 101                         | 172                     | 153                     |
| 10                                     | 144                         | 281                     | 259                     |
| 20                                     | 182                         | 420                     | 368                     |
| 50                                     | 270                         | 834                     | 491                     |
| 100                                    | 519                         | 1063                    | 681                     |
| 300                                    | 1262                        | 1627                    | 793                     |
| 500                                    | 2408                        | 1865                    | 870                     |
| 800                                    | 3854                        | 2216                    | 913                     |
| 1000                                   | 5701                        | 2642                    | 974                     |

After the relevant statistical data are selected, the data volume, traditional algorithm and parallel algorithm are compared. The experimental results of data mining are shown in Figure 1.

Figure 1. Performance comparison

Figure 1 shows that network big data mining based on computer technology framework is based on ignoring network communication overhead time. MR algorithm. However, when the volume of data is small, it has little difference from the single machine network big data mining algorithm in computing time. With the increase in the volume of data, its advantages are more prominent. Hence, when dealing with big data, the improved network big data mining algorithm can take advantage of the advantages of parallelism.

As an important standard to measure the performance difference between serial system and parallel system, the acceleration ratio is formulated as follows:

\[
SP = \frac{T1}{Tp}
\]  

(3)

SP is the acceleration ratio, T1 is the running time of single processor, TP is the running time of parallel system with P processors. The experimental results are shown in Figure 2.

Figure 2. SP acceleration ratio
Figure 2 shows that the network big data mining algorithm. When MR processes small data sets (such as 500000-1 million records), SP is less than 1, and the computation time of single machine network big data mining algorithm is less than that of the improved algorithm; however, with the increase of data volume, the computation speed of the improved algorithm is fast, SP increases; at the same time, with the increase of map node number, the increase rate of acceleration ratio will shrink, and tend to be flat. According to the experimental results, in data mining, we should first analyze the size of data scale, and then combine the size of data to decide how many node machines to call.

4. Conclusion
This paper mainly studies the optimization of network big data mining algorithm based on the traditional association rules of computer technology in the big data environment to effectively improve the computation speed and efficiency of the computation method in the improved data processing environment. Subsequently, the landslide data mining is taken as an example to apply the optimized algorithm to the big data mining system, and obtain the general rules of landslides, identify the frequent term set of landslide causes. Then, the candidate sets are filtered to define the landslide type. Finally, the closest or most approximate expression of landslides is obtained to provide a strong basis for reducing and preventing natural disasters.

Acknowledgments
This papers’ project is The 2018 basic research ability improvement project of young and middle-aged teachers in guangxi universities and colleges "Research on Information Resource Service Innovation of Rural E-commerce under Big Data Environment-Taking Hechi Region as an Example", Project Number: 2018KY1171.

References
[1] D'Alconzo, A. , Barlet-Ros, P. , Fukuda, K. , & Choffnes, D. . (2016). Machine learning, data mining and big data frameworks for network monitoring and troubleshooting. Computer Networks, 107(pt.1), 1-4.
[2] Njah, H. , Jamoussi, S. , & Mahdi, W. . (2018). Deep bayesian network architecture for big data mining. Concurrency and Computation Practice and Experience(1), e4418.
[3] Tan, Y. , Shi, Y. , & Tang, Q. . (2018). [lecture notes in computer science] data mining and big data volume 10943 a comparative study of network embedding based on matrix factorization. , 10.1007/978-3-319-93803-5(Chapter 9), 89-101.
[4] Finlay, & Steven. (2014). Predictive analytics, data mining and big data text mining and social network analysis. , 10.1057/9781137379283(Chapter 9), 179-193.
[5] Chu, & Wesley, W. . (2014). [studies in big data] data mining and knowledge discovery for big data volume 1 a generalized approach for social network integration and analysis with privacy preservation. , 10.1007/978-3-642-40837-3(Chapter 8), 259-280.
[6] Chen, B. S. . (2017). Big mechanisms in systems biology big tumorigenesis mechanisms in systems cancer biology via big database mining and network modeling. 431-526.