Design of a Computer Network Log Analysis System Based on Big Data Analytics Technology"

Andi Cui
Liaoning Qiming Auto Electrical Co., Ltd., Jinzhou, Liaoning, China

Abstract: The analysis of computer network logs is critical to network security. Traditional log analysis methods have difficulty identifying potential issues within massive amounts of data. This paper proposes the application of big data analysis techniques to computer network log analysis. A K-means algorithm based on a genetic algorithm is designed, and the simulation results show that the performance of the improved K-means algorithm has been enhanced. The computer network log analysis system designed using big data analysis technology adopts a six-layer architecture design. The system's functional design utilizes UML technology.

Keywords: Big Data, Computer Network Logs, Functional Architecture, Database Design.

1. Introduction

In the information age, the construction and application of informatization in enterprises and institutions are increasingly strengthened, accumulating a large amount of software and hardware resources. These software and hardware resources generate and record a large number of logs daily, which document the operation of computer network devices, network access conditions, etc. For system maintenance personnel, computer network logs are very important. Through the analysis of computer network logs, one can understand the operation of the system, such as historical access records, system anomalies, system failures, etc., laying the foundation for improving the management of computer software and hardware resources and strengthening network security. With the significant increase in computer network logs, traditional computer network log analysis methods can no longer meet the needs of log data growth[1-2]. As big data has become an important information technology tool in the operation of enterprises and institutions, the application of big data analysis techniques to computer network log analysis has also begun to receive attention. The application of big data analysis technology requires the help of system platforms to satisfy the collection, mining, and analysis of computer network log data, which is an advantage and condition not possessed in traditional computer network log analysis[3]. This paper designs a computer network log analysis system based on big data analysis technology, aiming to apply big data analysis to the analysis of massive computer network logs, change the traditional computer network log analysis model, improve the efficiency and level of computer network log analysis, and better play the role of computer network log analysis.

2. Overall Design of Computer Network Log Analysis System Based on Big Data Analysis Technology

In the computer network log analysis system based on big data analysis technology, logs are the main data collection method and source. According to different types of software and hardware, the data type and content of computer network logs are different, hence log data is characterized by diversity, massiveness, rapid generation, and high value [4-7]. Facing computer network logs with big data characteristics, to meet the needs of computer network log analysis, the paper constructs a computer network log system architecture diagram as shown in Figure 1.

![Figure 1: Architecture of Computer Network Log Analysis System Based on Big Data Analysis Technology](image)

The architecture of the computer network log analysis system mainly includes five levels: the Log Source Layer, Collection Layer, Storage Layer, Business Layer, and Display Layer. The Log Source Layer collects various types of software and hardware devices in the computer network, continuously providing a source of log data. The Collection Layer satisfies the collection of log data. The Storage Layer implements distributed storage of original logs, and the results of log analysis are also stored in this layer. The Business Layer fulfills the analysis requirements of log data. The Display Layer presents the processed and analyzed results of the log data for reference by system administrators and others.

3. Core Algorithm Design

Currently, traditional data analysis methods are used for computer network log analysis. Using traditional network log analysis methods, it is difficult to mine potential information from massive data. To solve this problem, this paper proposes using the K-means algorithm for network log information clustering analysis [8-10]. The core idea of the K-means algorithm is to divide the dataset M according to the provided
clustering center points. Data clustering division is evaluated based on the distance between the data points to assess their similarity. The closer the distance between the data, the higher their similarity. The K-means algorithm is as follows:

Step 1: Randomly obtain K cluster centers from the dataset M.
Step 2: Allocate the data points in the dataset to each cluster center using the Euclidean distance method, with the nearest distance as the center point.
Step 3: Replace each cluster center point and calculate the new cluster center point using the average method.
Step 4: Perform Steps 2 and 3 iteratively until the data in the cluster centers does not change or satisfies the set threshold.

The traditional K-means algorithm is greatly influenced by the cluster centers during its operation, which can easily lead to local optimal solutions. To solve this problem, it is proposed to incorporate a genetic algorithm into the traditional K-means algorithm. The improved algorithm flowchart can be seen in Figure 2.

![Flowchart of Improved K-means Algorithm](image)

**Figure 2: Improved K-means Algorithm Process.**

The improved K-means algorithm incorporates a genetic algorithm, using the variation operation of the genetic algorithm to optimize the cluster center location. The core steps of the improved algorithm are:

1. **Encoding:** Encoding is mainly used for the cluster number k. The value of k is $k=\lceil\sqrt{n}\rceil$, an integer between a certain range. In this paper, binary encoding rules are used for encoding. The chromosome length is set to 8 bits. The first 7 bits use binary encoding rules, and the last bit corresponds to individual fitness. (2) **Population initialization:** In the improved K-means algorithm, the number of chromosomes in the population is controlled at about 100, the crossover probability is controlled between 0.8 and 0.9, and the mutation probability is controlled between 0.001 and 0.04. (3) **Clustering:** Decode each chromosome in the population to get the category number k. (4) **Fitness function:** The improved algorithm uses the following formula as the fitness function.

$$f(P_i) = \sum_{i=1}^{k} \sum_{C_j} | P - m_j |$$  \hspace{1cm} (1)

In formula (1), $P$ represents the clustering data and $m_i$ represents the clustering center point.

(5) Genetic operations, the improved algorithm uses the roulette gambling algorithm for genetic operations. (6) The algorithm ends. When the fitness value in the improved algorithm does not change, the algorithm stops; otherwise, repeat the operations (3)-(5).

To evaluate the performance of the improved algorithm, this test uses Iris data and glass data, with a test data volume of 200, the data attribute is the optimal number of clusters 3, and each data has 4 attributes. The test results are shown in Table 1.

<table>
<thead>
<tr>
<th>Algorithm Type</th>
<th>Number of Iterations</th>
<th>Accuracy Rate (%)</th>
<th>Number of Iterations</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional K-means Algorithm</td>
<td>45</td>
<td>72.34</td>
<td>38</td>
<td>76.54</td>
</tr>
<tr>
<td>Improved K-means Algorithm</td>
<td>31</td>
<td>83.45</td>
<td>29</td>
<td>90.42</td>
</tr>
</tbody>
</table>

It can be seen from Table 1 that the correct rate of the improved K-means algorithm is obviously higher than that of the traditional K-means algorithm.

4. Design of Computer Network Log Analysis System Using Big Data Analysis Technology

4.1 System Requirements Analysis

The designed computer network log analysis system using big data analysis technology should meet the following requirements:

1. **Dynamic Data Reception and Storage:** The system should be able to dynamically receive and store log data.

2. **Scalability and Storage:** The designed log analysis system should be capable of storing massive data, with good scalability. As the storage volume increases, the system's performance should not be significantly affected.

3. **Fast Response Time:** The system should have fast response times to ensure timely analysis. Quick response is crucial for achieving the desired analysis objectives.

4. **User-Friendly Interface:** The system interface should be simple and intuitive. Considering the wide age range of users and the goal of analyzing computer network logs, the system should be designed to be user-friendly.

5. **Data Localization and Segmentation:** The system should enable data localization and segmentation to meet user query requirements. Users should be able to locate and segment data effectively for their specific queries.
4.2 System Architecture Design

The designed computer network log analysis system using big data analysis technology adopts a 6-layer architecture, as shown in Figure 3.

The top layer is the visualization layer, responsible for managing the computer network log analysis functionalities such as user login, data analysis, and data collection.

The business layer manages the business logic analysis of computer network log analysis using big data analysis technology. It utilizes algorithms provided by the big data module and incorporates the improved K-means algorithm designed in this paper into the business logic layer. This layer includes Mapper and Reducer, which serve as input and output layers. The input layer typically stores network log information in the Hadoop distributed system, utilizing Hadoop for file path retrieval. The first Map is completed, followed by more map tasks. The output of the map tasks is then passed to the reducer layer. The improved K-means algorithm is placed in the MapReduce layer, enabling disk data reading and deep mining of network log information.

The log layer mainly obtains the network log data source. The data sources for generating logs in the log layer mainly include firewalls, routers, switches, intrusion detection devices, etc.

The log analysis module is designed using UML modeling technology, with the main class being Loganalysis. This class contains the classes Datacleaning, K-means, and Resultdisplay, which are used for data cleaning management, data analysis management, and result analysis management respectively. The Datacleaning class contains the methods...
readdata(), codeture(), and savedata(), which are used for data reading, data encoding judgement, and data saving, with a return type of boolean. The K-means class contains the methods normalization() and cluster(), which are used for normalization processing and clustering processing. The Resultdisplay class contains the methods SaveResults() and saveShow(), which are used for result saving and result displaying.

(2) Data Collection Module. The data collection module contains the functionalities of data collection and data storage. The data collection module contains the classes: dataacquisition, dataacquisitionBLL, dataacquisitionDALL, and dataacquisitionMobile, which are used for data collection interface management, data collection business logic management, data collection data management, and data collection dataset management respectively. The Dataacquisition class contains the methods: dataacquisitionfun() and dataacquisitionsave(), which are used for data collection and data storage management respectively.

5. Conclusion

With the increase in the number of computer network logs, the demand for log analysis continues to rise. The application of big data analysis technology to the design of computer network log analysis systems effectively enhances the efficiency of log data mining and analysis. It provides a reference for enterprises and institutions to strengthen computer network log analysis, improve computer network equipment fault management, and ensure network security.

References