K-means Clustering Optimization Algorithm Based on MapReduce

Zhihua Li\textsuperscript{1,a}, Xudong Song\textsuperscript{2,b}, Wenhui Zhu\textsuperscript{3,c}, Yanxia Chen\textsuperscript{4,d} *  
\textsuperscript{1}College of Network Engineering, Shijiazhuang Institute of Technology, Shijiazhuang, 050228, China  
\textsuperscript{2}Software Institute, Dalian Jiaotong University, Dalian, Liaoning, 116028, China  
\textsuperscript{3}Beijing Datang Telecom convergence communications Technology Co., Ltd., Beijing, 100029, China  
\textsuperscript{4}Department of Medical Imaging, Dalian Medical University, Dalian, Liaoning, 116044, China  
\textsuperscript{a}email: intand@163.com, \textsuperscript{b}email: xudongsong@126.com, \textsuperscript{c}email: 553955162@qq.com, \textsuperscript{d}email: cyx_dl@126.com, * Corresponding author

Abstract

Aiming at the defects of traditional K-means clustering algorithm for big data, this paper provides K-means clustering mining optimization algorithm based on big data, shows a MapReduce software architecture which is suitable for large data processing mechanism, provides an improved method for selecting initial clustering centers and puts forward a K-means algorithm optimization based on MapReduce model. The improved algorithm is applied to the coal quality analysis, the result shows that compared with traditional algorithms, the optimization algorithm improves the efficiency of the algorithm obviously, and the accuracy is also enhanced.

Keywords: Data Mining, K-means Clustering algorithm, MapReduce, Hadoop

1 Introduction

K-means clustering algorithm is a classical clustering algorithm based on splitting method. Because the theory of the algorithm is reliable, simple and convergent rapidly, K-means algorithm is widely used \cite{1,2,3,4,5}.

However, with the development of the information society, the data size the data mining tasks faced is more and more big. Even though the traditional clustering mining optimization algorithm have good accuracy in the face of massive data, its time complexity of serial calculation method are high. More how to store, handle these massive amounts of data, and dig out further useful knowledge can guide the application become a thorny issue. Aiming at the defects of traditional algorithm, this article proposes an improved method for selecting initial clustering centers and puts forward a K-means algorithm optimization based on Hadoop cloud computing platform. The improved
algorithm is applied to the coal quality analysis, the results show that compared with traditional algorithm, the optimization algorithm improves the efficiency of the algorithm obviously, and the accuracy is also enhanced.

2 Operation mechanism of MapReduce

Hadoop is an open source distributed computing platform, which mainly consists of distributed computing framework--MapReduce and distributed file systems--HDFS. MapReduce is one of the core components of Hadoop, and it is easy to realize distributed computer programming by MapReduce on Hadoop platform.

MapReduce is a software framework for parallel computing programming model of large-scale data sets, having obvious advantages in dealing with the huge amount of data.

Operation mechanism of MapReduce is as follows:
(1)Input: MapReduce framework based on Hadoop requires a pair of Map and Reduce functions implementing the appropriate interface or abstract class, and should also be specified the input and output location and other operating parameters. In this stage, the large data in the input directory will be divided into several independent data blocks for the Map function of parallel processing [6][7].

(2)MapReduce framework puts the application of the input as a set of key-value pairs <key,value>. In the Map stage, the framework will call the user-defined Map function to process each key-value pair <key,value>, while generating a new batch of middle key-value pairs <key,value>.

(3)Shuffle: In order to ensure that the input of Reduce outputted by Map have been sorted, in the Shuffle stage, the framework uses HTTP to get associated key-value pairs <key,value> Map outputs for each Reduce; MapReduce framework groups the input of the Reduce phase according to the key value.

(4)Reduce: This phase will traverse the intermediate data for each unique key, and execute user-defined Reduce function. The input parameter is <key, {a list of values}> , the output is the new key-value pairs <key, value>.

(5)Output: This stage will write the results of the Reduce to the specified output directory location.

Operation mechanism of MapReduce is shown in Figure 1.
3 Clustering mining optimization algorithm based on MapReduce

The data set processed by MapReduce should have such characteristics: It can be broken down into many small data sets, and each small data set can be completely parallel processed [8][9][10]. The process of K-means algorithm based on Hadoop mainly has two parts, the first part is to initial clustering centers, and divide the sample data set into a certain size of data blocks for parallel processing. The second part is to start the Map and Reduce tasks for parallel processing of algorithm in time, until process gets the clustering results. Its algorithm process is shown in Figure 2.

---

The initial clustering centers of traditional algorithm selected randomly, will
cause the instability of clustering results. This paper adopts a method of the initial clustering center selection to improve the stability of the results. Optimized K-means clustering algorithm firstly choose k samples to initialize clustering centers according to certain algorithmic rules, then k clustering centers are stored in a file on the HDFS as a global variable [11].

Let cluster sample data set: \( D = \{d_i | d_i \in \mathbb{R}, i = 1, 2, 3, \ldots, n\} \), k cluster centers are showed by \( c_1, c_2, c_3, \ldots, c_k \). Specifically definitions are as follows:

1. In the data set, distance between any two n-dimensional vector is expressed using Euclidean distance:
   \[
   \text{dist}(d_i, d_j) = \sqrt{(d_{i1} - d_{j1})^2 + (d_{i2} - d_{j2})^2 + \cdots + (d_{in} - d_{jn})^2}
   \]

2. Data center of sample points \( O(d_i, d_j) = \frac{(d_{i1} + d_{j1})}{2}, \frac{(d_{i2} + d_{j2})}{2}, \ldots, \frac{(d_{in} + d_{jn})}{2} \)

3. The average distance between sample points: \( \text{averg} = \frac{\sum \text{dist}(d_i, d_j)}{c_k} \)

4. Experiment and Result Analysis

Ourexperimental data set is from a coal group enterprise. This experiment analyzes results respectively from the effectiveness of the algorithm and speed ratio, using K-means optimization algorithm based on Hadoop for completing the clustering analysis of characteristic data of the coal group enterprise.

We use amachine as the NameNode and JobTracter node; five other machines are DataNode and TaskTracker node. Each node hardware configuration is as follows:

- CPU is i5 M 480 @ 2.67 GHz dual-core, memory is 1 g. Hard disk is 250 g / 7200 RPM.
- There are 18,038 coal experimental data sample points, using the traditional K-means algorithm and optimization algorithms to test to generate four clusters. Traditional clustering algorithms due to the dependence of the initial cluster centers lead to instability of clustering results, which clustering results that different experiments produced are constantly changing, and results of
optimization algorithms remain unchanged. This paper selected two traditional clustering algorithms of results and the optimization clustering results shown in Figure 3, Figure 4, and Figure 5.

5 Conclusion

Aiming at the defects of traditional K-means clustering algorithm for big data, this article improved the selection of the initial clustering center firstly, secondly article realized the parallelization of K-means algorithm using operation mechanism of MapReduce. Experiments show that the improved algorithm has better effectiveness and higher computational efficiency compared with the traditional algorithm and the greater the amount of data the more obvious advantages.

Acknowledgment

This work has been supported by China Scholarship Council and National Natural Science Foundation of China (No. 61074029).

References


