Application of Parallel Clustering Algorithm Based on Domain Decomposition

Xue Tian

Software Engineering Department, Inner Mongolia Electronic Information Vocational Technical College, Hohhot 010010, China

Abstract

In recent years, with the rapid development of computer technology and the popularization and promotion of the Internet, there has been a data growth of geometrical progression, which indicates that China has ushered in a big data era. However, the increasing number of data has also brought many technical problems, and how to excavate and analyze huge data has become an urgent problem to be solved. Today, cloud computing technology has become a hot topic of research; Hadoop is an open source architecture platform in cloud computing, which has unparalleled advantages in parallel computing of massive data, but the traditional clustering algorithm adopted by Hadoop platform still follows the equal data allocation method, which will undoubtedly have a serious impact on the performance of MapReduce. To this end, this paper proposed a parallel clustering algorithm based on domain decomposition, namely DBSCAN parallel clustering algorithm, and an in-depth research on the data distribution method and parallelization of the clustering algorithm was conducted, based on which the application of DBSCAN clustering algorithm based on domain decomposition was studied.

Keywords: Domain Decomposition, Parallel Clustering Algorithm, Hadoop Platform, Cloud Computing, Proportional Data Allocation.

1. INTRODUCTION

1.1 Research Overview

The advent of the era of large data enhances the increase of the number of data moment by moment, which makes it difficult for the traditional simple summary and analysis method to explore and analyze the valuable information in the massive data, and it is urgent to solve these problems in a more efficient way, so as to help people to use these valuable information for decision-making analysis. With the rapid development of science and technology, many advanced concepts and methods such as statistics, machine learning and pattern recognition have been applied in the field of data mining, and data mining technology has emerged. Nowadays, people pay more and more attention to data mining technology, and the application of data mining technology can enable people to excavate and analyze the latent knowledge in the data in a better way, so as to assist people in making more scientific and accurate decisions. In large data clustering, the clustering algorithm based on density is the most typical and traditional clustering algorithm, which can only be applied to small-scale data input, and it is difficult to meet the demands of large high dimensional data processing (Li and Li, 2017). The first phase of the clustering algorithm is PDBSCAN algorithm, which can effectively reduce the number of domain queries, make the I/O cost significantly lower, and then achieve rapid clustering, but such phenomenon as partial incomplete clustering may appear. Some scholars have improved the algorithm to achieve data division based on the distribution characteristics of mass data before the clustering, which improves the efficiency of clustering algorithm and the memory cost is further reduced; however, this clustering algorithm is more complex and also lower in parallel efficiency (Liu et al., 2017). In this paper, a clustering algorithm based on domain decomposition is deployed in the cloud computing Hadoop open source platform to solve the clustering problem of massive data, namely DBSCAN clustering algorithm. Because the data itself and the Hadoop environment of different structures contain noises, the data clustering algorithm cannot be directly transplanted; taking into account the parallelization characteristics of the clustering algorithm, the DBSCAN clustering algorithm and the cloud computing parallel mode can be combined to realize the parallelization processing of massive data by DSBCAN clustering algorithm.
1.2 Research Purposes

The aim of this paper is to propose a DSBCAN parallel clustering algorithm based on domain decomposition for the clustering problem of massive data and to realize the clustering processing of massive data by DSBCAN parallel clustering algorithm by combining the DSBCAN parallel clustering algorithm with cloud computing technology. In this paper, the concept of clustering and the steps of clustering algorithm were expounded, the concept of DBSCAN algorithm was introduced in detail, the related technology of concurrent clustering was introduced briefly, and the proportional data allocation method was proposed for the Hadoop platforms and data of different structures containing noises. In addition, the parallel processing strategy of DBSCAN clustering algorithm based on domain decomposition was designed, and the MapReduce implementation steps of the DBSCAN parallel clustering algorithm were clarified, based on which the practical application of the method of proportional data allocation and the DBSCAN parallel clustering algorithm were experimentally tested.

2. SUMMARIZATION OF PARALLEL CLUSTERING ALGORITHM

2.1 Clustering Algorithm

The development of modern information technology has promoted the development of data mining technology, and people hope to find valuable information through data mining technology to make forward-looking decisions. Data mining technology is mainly divided into two types, namely, descriptive and predictive; cluster analysis, association analysis and some other methods are of the predictive kind and the cluster analysis has become the most popular method. The so-called clustering algorithm is to carry out decomposition on the big data domain to enable the similar objects in the domain to be gathered in the same class in accordance with the agreed measurement method, so as to clear the data distribution and conduct more in-depth analysis of the specific class of the big data, such as the nearest neighbor algorithm, DBSCAN and K-means, which are all very classical clustering algorithm (Gao et al., 2015). Traditional data analysis methods can only deal with numerical data effectively, but today's data types are diverse, which makes it difficult for traditional data classification to meet the needs of today's analysis; in this situation, the clustering algorithm has been constantly improved and changed, and the diversification of data types makes the distance measurement method very different. Therefore, in order to improve the clustering efficiency, it is necessary to synthetically consider the characteristics of various types of data when adopting the clustering algorithm (Du et al., 2015). Clustering algorithm can be divided into the following steps in classifying data: first, to preprocess the data, in which the original input data are formatted and regularized, the anomaly points are detected and cleared, and the appropriate features are selected and extracted. Second, to cluster the data, analyze the characteristic vector similarity of the data by Euclidean distance, and cluster the data using the corresponding clustering algorithm. Third, to evaluate and describe the results of clustering and convert the results into valuable information (Zhou et al., 2014).

2.2 Dbscan Parallel Clustering Algorithm

The DBSCAN parallel clustering algorithm conducts decomposition and clustering of the domain based on each data object and specific value in the cluster. The object in the EPs field must exceed the given value, and when the number of data points or objects in the adjacent domain is found to be more than or equal to the given value, the cluster will continue to be able to filter the noise data effectively. It uses distance, density and other parameters to evaluate the object's similarity; the domain density refers to the number of data objects in the unit area; the definition of the unit area is measured by distance, and the commonly used metric distance includes Manhattan distance and Euclidean distance (Ding et al., 2016). DBSCAN clustering algorithm has four important definitions, namely, domain, core object, boundary object and noise data.

2.3 Introduction to the Related Technologies of Concurrent Clustering

Parallel computing is the allocation of computational tasks so that multiple processors can perform computational tasks in parallel, making processing faster. The parallel computing method can significantly improve the processing and computing efficiency; it will decompose the dataset into several sub datasets, and the data sets can be solved by the cooperation of multiple computers. There are four main structures, namely, SMP (symmetric multi-processing) structure, DSM (distributed shared storage multi-processing) structure, MPP (massive parallel processing) structure and cluster structure (Wu et al., 2016). In parallel clustering, the most commonly used techniques include sequential method, partition method, sampling method, data summarization method and parallel method.
3. RESEARCH OF DBSCAN PARALLEL CLUSTERING ALGORITHM IN HADOOP PLATFORM BASED ON DOMAIN DECOMPOSITION

3.1 Proportional Data Allocation Method

In Hadoop platforms of different structures, the disk capacity and resource allocation of each node are quite different, which makes the computing ability of the platform different; if each node in the platform structure is allocated in the same amount of data, it will inevitably reduce the performance of the system; therefore, to improve the performance of the platform, a method of allocating proportional data needs to be proposed (Yang and Zhao, 2015). To this end, this paper presents the steps for the realization of the proportional data distribution method. First, as for the data allocation algorithm in the initial Hadoop platform, the data partitioning method is used to conduct domain decomposition of the user input data in the Hadoop platforms of the same structure to separate them into several files of the same size. Second, conduct calculation of the relative processing rates of the node, as in the Hadoop platforms of different structures, the task processing speed of each node varies greatly, which makes it necessary to define the node completion with the definition formula

\[ J_{node} = \alpha \frac{V_{disk}}{maxV_{disk}} + \beta \frac{V_{cpu}}{maxV_{cpu}} \]

where \( J_{node} \) represents the completion of nodes, \( \alpha \) and \( \beta \) represent parameters, which are very hard to determine and require experimental tests and statistics of the completion time of each node through the different application degree to obtain the relative processing rate of the node. To make the statistics of the processing speed, all nodes need to be allocated with the same data, and then the processing time of the node needs to be recorded, among which the longest processing time is taken as the reference value and the relative calculation rate of other nodes; the higher the value of the node relative calculation rate is, the faster the processing speed of the representative node will be. Third, sort all the calculated relative computation rates of nodes according to their values, and divide the total number of the virtual slots using \[ V = \sum_{i=1}^{m} (\text{int}) \frac{J_i}{J_m} + 1 \]

where \( J_i \) and \( J_m \) represent the i node and the slowest relative computation rate; the number of virtual slots is represented by \( V \) and the value of \( J_i/J_m \) is represented by \((\text{int})\), by which it can be obtained that there are \( N/V \) file fragments in each virtual slot. Four, calculate the number of fragments in all nodes, sort them by serial number, and map the virtual slots of these nodes to the map task slot. Five, allocate the data proportionally, and the master is responsible for storing the computing rate of each node and the number of allocated blocks, etc. The storage is conducted in the form of a list, and the input file block is allocated by Jobtracker according to the rate of calculation for each node in the above four steps (Ma and Yang, 2015). Figure 1 shows a schematic diagram of the proportional data allocation method.

![Figure 1. Schematic Diagram of Data Distribution Method](image)

3.2 Parallelization Research of Dbscan Clustering Algorithm

DBSCAN clustering algorithm based on domain decomposition is able to effectively handle noise data sets, and realize the parallel processing of Hadoop platforms of different structures by DBSCAN clustering algorithm through the introduction of the parallelization design idea (Zhu et al., 2017). The parallel design is mainly divided into the following steps, first, it is needed to conduct domain decomposition of large data sets, recombine the data blocks in the several decomposed domains in accordance with the above proportional data allocation method, and allocate the data tasks according to the computing proportion of different computers; second, apply DBSCAN clustering algorithm to each child node of data blocks; third, combine the calculation results of each computer; fourth, rename the clustering results of the DBSCAN clustering algorithm; five, output the renamed clustering results (Wang et al., 2016). Figure 2 shows the parallel frame diagram of DBSCAN clustering algorithm.
Previously, the parallel design idea of DBSCAN clustering algorithm was studied; however, when the input database was decomposed into a number of identical and disconnected domains in the process of file storage in the Hadoop platform, the size was set according to the user, and the decomposition domain method was still used in the data clustering, which would turn some core points into noise points; because the boundary areas were not connected, the data objects’ internal connection was cut off, resulting in the clustering result deviation. Therefore, it is necessary to introduce the concept of cross domain to solve the above problem, so that it can be used in the domain decomposition of the database to include the adjacent partition boundary and it only needs to scan the boundary area to improve the efficiency of the algorithm (Jiang et al., 2016). The Dbscan clustering algorithm is divided into four stages in the parallel processing of Hadoop platforms of different structures, namely, Map stage, Reduce stage, merging result stage and label clustering result renaming stage. In the Map stage, the local data domain is constructed for KD-Tree spatial index. Then, the DBSCAN algorithm is executed in the decomposed domain, and it is decomposed into boundary domain and local domain; the result of the boundary area is stored by HDFS, and the local domain result is stored by the local disk. The mark for judging the data point as the core point is iscore. In the result of DBSCAN Clustering algorithm, CID represents the number of cluster identities. In the Reduce phase, all existing same index data points in different domains are summarized; if the data point is a core point in all parts, then the judgment mark is true, which means that the node is at least a core point in an area; identifying the core point mark can help to merge the DBSCAN clustering algorithm. In the merge results stage, if the number of CID nodes is more than one, and these CID nodes are all core points, they can be merged and then renamed. In the clustering results renaming stage, the clustering results need to be renamed to ensure that there is only one global name for each cluster (Li et al., 2017).

4. APPLICATION RESEARCH OF DBSCAN PARALLEL CLUSTERING ALGORITHM BASED ON DOMAIN DECOMPOSITION

4.1 Research on Application of Proportional Data Allocation Method

In order to study the application of proportional data allocation method, it is needed to construct the application environment of Hadoop platforms of different structures, which is composed of five PCs, with the master control node master controlled by 1 PC and the Worker working node controlled by 4 other PCs. The operating system adopts ubuntu10.04, setting the directory structure in all nodes consistent with the username, and the username is all set to root. Hadoop is written in the Java language, and it should be noted that SSH should be configured before configuring Hadoop. The application performance of the proportional data allocation method in different structure Hadoop is tested by using statistical words and Sort to facilitate comparison. 2.5GB and 1.5GB input files are used to run WordCount and Sort applications in 4 working nodes of different structures, and the test results show that the size change of the input file does not affect the proportional data allocation method; however, when the input file size is consistent, it can be seen that the WordCount distribution time is longer, which can prove that different Applications have different calculation rates in each node. In the test of different data allocation algorithm, Sort application relative velocity distribution method, WordCount relative rate allocation method, equal data distribution method in the homogeneous environment, and the method of all input data stored in the control node are adopted to conduct the comparison test for the application effect of the proportion data distribution method, and the comparison results show that the allocation mechanism of the former two methods of data distribution proportion in different Hadoop clusters are significantly better than that of the latter two data distribution methods (Li and Qin, 2016).

4.2 Application Research of Dbscan Parallel Clustering Algorithm Based on Domain Decomposition

To analyze the application effect of DBSCAN parallel clustering algorithm based on domain decomposition, a
group of real data sets and four groups of combined data sets are used for validation; the combined data sets are nt8, 800k, nt7, 1000k, nt5, 800k, nt4, 800k, and the real data set is the machine learning data set. From the test of application effect of execution time by the Griddbscan algorithm and the DBS algorithm, it is known that the execution time of the DBSCAN algorithm is smaller than that of the GRIDDBSCAN algorithm when the same data set is executed, because the DBSCAN uses the cross data domain decomposition, which saves a lot of time in the eventual merging of clustering results. In order to analyze the scalability of DBSCAN algorithm application, it is needed to duplicate the real data set according to the characteristics of data for many times to generate data sets of different sizes; the data results show that when the data volume is small, the operation efficiency of DBSCAN parallel clustering algorithm based on domain decomposition is much lower than that of single computer, but the problem can be solved through the expansion of the nodes. When the amount of data is large, the processing time of single computer is much more than that of DBSCAN parallel clustering algorithm (Yan and liu, 2016). Therefore, the processing efficiency of DBSCAN parallel clustering algorithm based on domain decomposition is much higher than that of single computer, and it is more suitable for parallel processing of massive data.

5. CONCLUSION

To sum up, by means of the method of proportional data allocation, the serious influence of the traditional equal data distribution method on the MapReduce performance can be effectively reduced, and at the same time, by incorporating the concept of cross data domain decomposition in the DBSCAN algorithm, the problem of cutting down the internal connection between data objects in data partitioning can also be effectively solved, so as to realize the parallel processing of DBSCAN clustering algorithm based on domain decomposition in Hadoop platforms of different structures; compared with other clustering algorithms, the DBSCAN clustering algorithm based on domain decomposition is more suitable for the parallel processing and analysis of massive data, which greatly improves the operation efficiency of the system.

REFERENCES

Yang Y., Zhao C.G. (2015). Research and implementation of parallel approximate spectral clustering algorithm based on Hadoop MapReduce, Computer applications and software, 32(8), 17-21+63.