Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

Data Mining Method in Forest Engineering Based on Big Data and Internet of Things Technology

Chen Ji-chao[1*]
[1]Sanquan College of Xinxiang Medical University, Xinxiang, Henan, China
*Corresponding Author.

Abstract:

A large amount of forest data in forestry engineering leads to that data management is a very complex work. Data mining mainly includes four parts: clustering analysis, prediction modeling, association analysis and anomaly detection. In order to get the potential valuable information from the complex data, it is necessary to deeply study and flexibly use the data mining algorithm. Based on spark distributed framework, this paper focuses on the maximum frequent itemset mining algorithm and density clustering mining algorithm. In the aspect of frequent itemset mining, because of its high value, advanced information is hidden in long frequent items. Therefore, mining maximal frequent itemsets has higher value. After combining the advantages of the existing algorithms, a recursive deep path search is proposed to generate the maximum frequent item candidate set at one time. Then, the candidate frequent itemsets are sorted by length first. Then the improved process of superset test is recycled. The experimental results show that the improved algorithm optimizes the pruning and dimensionality reduction of the data set, and reduces the scale and mining times of the candidate item set. This method solves the problem of low efficiency of the existing maximum frequent mining algorithm when the amount of data is large and the dimension is high.

Keywords*: Forest engineering, data mining, clustering analysis, anomaly detection, density clustering mining algorithm.*

I. INTRODUCTION

In recent years, information technology has developed rapidly, and the data volume has been explosive growth and accumulation [1-2]. In the fields of e-commerce, Internet, scientific research, finance and other applications, a large number of data will be produced every day, especially with the rise of Internet of things and social network technology, the rapid growth

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

and accumulation of data, which makes people pay more and more attention to big data[3]. The characteristics of big data are volume, variety, velocity, variability and value. In the actual production and life, people want to dig out the more valuable information in big data.

In order to extract valuable information from large and complex data, it is necessary to know the relevant fields and the corresponding data mining algorithms. In the aspect of processing small data, data mining algorithm has been relatively mature. Some classical algorithms, such as clustering algorithm, classification algorithm, association rule algorithm, etc., have been widely used in practice [4-5]. Therefore, people urgently need to study the data mining algorithm for big data, which can efficiently process massive data. However, the data mining algorithms which run on single server and can only be used for those small data sets can not support the data volume which is growing in geometric data. The problem of low efficiency and slow speed caused by the increase of calculation has become the bottleneck of traditional data mining algorithm.

At present, after decades of research and development, data mining has been very mature, resulting in a variety of data mining algorithms. For example, K-means algorithm and fuzzy clustering algorithm in clustering analysis, decision tree algorithm and support vector machine in classification algorithm, Apriori algorithm in association analysis, etc. these algorithms and their improved algorithms have been applied in practice to provide decision support for people, especially in dealing with small-scale data [6-8]. Based on the background of distributed computing, through the study of the traditional classic data mining algorithm, this paper improves and proposes an efficient distributed data mining algorithm suitable for big data, so as to mine the potential value of data more effectively.

II. OVERVIEW OF DISTRIBUTED PLATFORM

Distributed parallel computing is to divide large-scale processing tasks and distribute them to different nodes and run at the same time, and then all the results are merged together. This is also the core of cloud computing.

2.1 Overview of Hadoop platform

Apache Hadoop implements the computing framework of MapReduce through the concepts of mapping and specification, and also provides a distributed storage system for large-scale data. The core idea is to map a set of key value pairs into a new set of key value pairs, and then send the key containing the same function value to the same reduce end.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

Because Hadoop is designed to deal with the large-scale available data in the Internet, it provides transparent data management and task management for users. Therefore, this transparent management mode enables the framework to support the cluster operation including thousands of units, and process Pb level data at the same time. The data in Hadoop is transmitted through the network and stored on the disk in the form of data block, which ensures the reliability of the data in the form of data block copy [9]. The framework of MapReduce is shown in Figure 1. In recent years, Hadoop has been widely used in cloud computing, cloud storage and big data processing. Well known Internet enterprises such as Yahoo, Facebook, Alibaba and Baidu have also applied Hadoop platform, which improves efficiency and increases enterprise benefits [10].

Although Hadoop has brought great convenience and advantages to Internet big data processing, there are still some problems to be solved and improved. These problems include the delay of data transmission in shuffle phase, repeated merging, and low disk access speed. First of all, the data input source and result of Hadoop cluster are stored on the disk of datenode node in the form of data blocks, and the data between Map and Reduce is transmitted through the network. Although with the development of technology, the network transmission rate has been greatly improved, but the huge increase in the amount of data, and the low efficiency of disk read-write rate, make the waiting time before the start of the reduce phase is still very long.

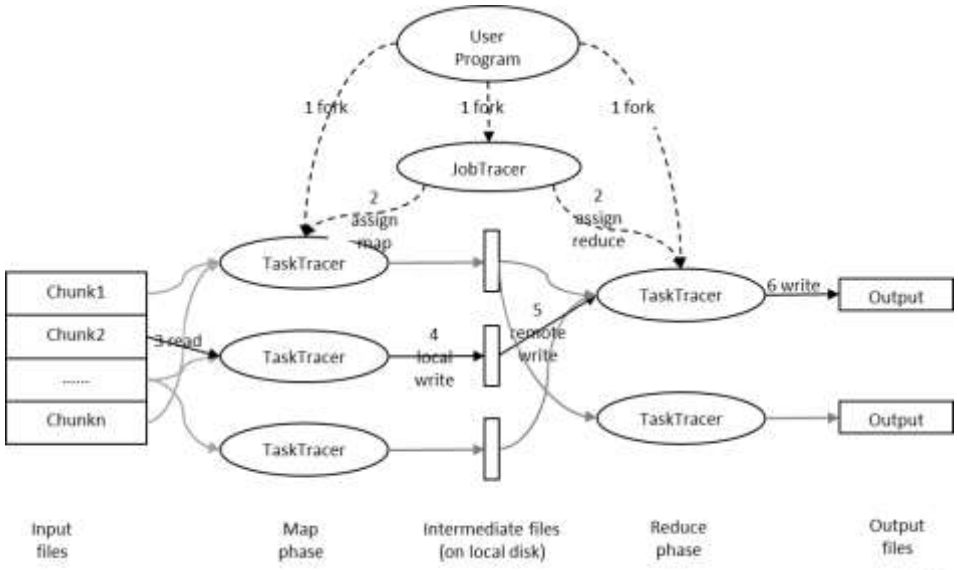

Fig 1: MapReduce framework diagram

2.2 Overview of Spark platform

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

In recent years, Spark platform, with its superior performance and rich development stack, has quickly become popular among major companies around the world. Compared with hadoop, Spark not only implements MapReduce programming model, but also abstracts distributed data into elastic distributed data set (RDD), and provides richer operators such as filter, join, groupByKey, etc., and provides corresponding inter-group for load balancing, remote communication and task scheduling, as shown in Figure 2. Its programming style is easy to understand, the bottom layer is transparent to users, and the upper layer provides users with a functional programming interface similar to Scala.

Compared with Hadoop MapReduce, Spark parallel computing framework is not only based on RDD, but also compatible with distributed storage layers such as HDFS and Hive. MapReduce outputs the results to the disk, which requires frequent reading and writing of the disk, and is inefficient. Spark not only abstracts the data set and stores it in memory. Moreover, the task execution is transformed into directed acyclic graph (DAG), and the stages are divided according to different types of operators. According to different stages, the program is not executed line by line and sentence by sentence. Instead, a series of RDD conversions are executed at one time in a delayed execution mode, which improves the efficiency and can trigger a recalculation for the wrong stage to achieve the purpose of fault tolerance. In terms of the overhead of task scheduling, hadoop has a very high delay in submitting a task in some extreme cases, while Spark adopts an event-driven mechanism to avoid the overhead of process or thread switching by reusing threads through a thread pool.

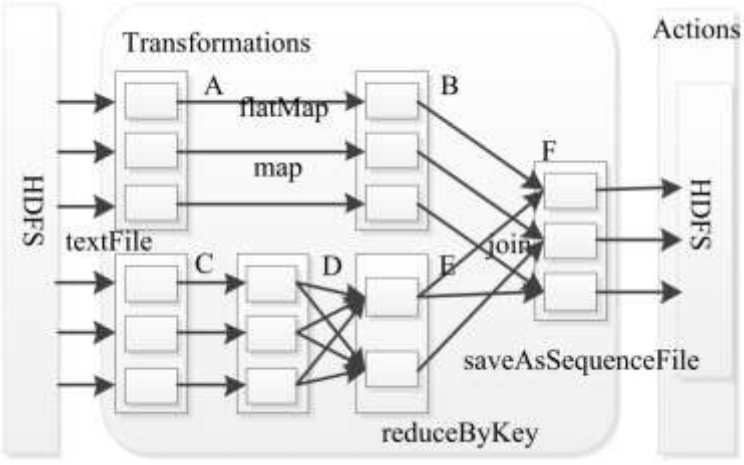

Fig 2: Spark programming model

III. RESEARCH ON DISTRIBUTED MAXIMUM FREQUENT ITEM ALGORITHM

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

3.1 Apriori algorithm and FP-Growth algorithm

Apriori algorithm is a prior algorithm, which uses a prior property, that is, "all non-empty subsets of frequent itemsets must also be frequent", and adopts the iterative method of layer-by-layer search, and every layer of exploration needs the results of the previous layer. The database needs to be scanned many times before generating the complete set of frequent patterns, and a large number of candidate frequent sets are generated at the same time, which makes Apriori algorithm time and space more complex.

In order to cope with the inability of Apriori algorithm in dealing with high-dimensional frequent itemsets, Jiawei Han proposed FP-Growth algorithm, which not only reduced the memory occupation, but also kept the relationships among attributes in the transactions by transforming the transactions in the dataset into tree structure. Firstly, the algorithm scans the transaction set D, finds out the frequent itemset and filters the data set with it. Then, taking "Null" as the root node, it builds FP-Tree, and adds every transaction in the data set to the frequent itemtree. At the same time, the following item header table mapping is used to record the frequency of each frequent item in the table and the pointer to its corresponding node in FP-tree. Finally, the FP-tree is recursively mined to find out all frequent item sets.

The FP-tree data structure is as follows:
FPTree{
Node: root; //root node
Map[item,Summary] :Summaries; //Project header table mapping
}
Node{
int: item;
int: count;
node: child,
node: parent
}
Summary{
int：count; //Node count of different path positions of the same Item
}
List[Node]: nodes; //different path location nodes of the same Item

3.2 Improved distributed SMFI algorithm

This paper puts forward the following plans:

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

(1) When the size of the data set is very large, the single machine environment can not accommodate it. Therefore, the partition projection method is adopted, and the partition number of an item is calculated by the partitioner, so that the records are divided into different sub records and put into different partitions, and the candidate subsets are formed in the parallel partitions of different nodes.

(2) In view of the high dimension of data itemsets, if we simply mine the frequent pattern tree and generate the candidate set of completely frequent itemsets, the higher the dimension of records, the more subsets it contains, so the more memory space the candidate set takes up. Therefore, for the frequent item tree of each partition, this paper uses the bottom-up method to carry out deep path recursive search on the filtered frequent item tree spanning tree, and finds out the longest path record as the indirect candidate frequent item set in different partitions that have been filtered. In this step, most of the invalid short sub record itemsets are removed, and the size of the candidate itemsets is greatly reduced. Finally, a cyclic superset test is performed on the indirect candidate frequent itemsets.

(3) In the process of screening the indirect candidate frequent itemsets generated in step 2, most existing algorithms often maintain an additional result set of the largest frequent itemsets, and perform superset test on each indirect candidate frequent itemset. This paper does not use the above method, but uses the method of first sorting by length and then ascending by prefix sequence number, and then carries out the superset test on the maximum frequent candidate set from top to bottom, from short to long. If the candidate record has no superset in the result set, it will be added to the maximum frequent result set.

3.3 SMFI algorithm description

The core of Spark is distributed elastic data set RDD, and the data of each partition exists in the memory of different machines. It provides some operable API interface functions, such as mappartion, map, reduceByKey, filter, etc., which respectively operate RDD at different levels. For example, map performs the same function operation on the data in each partition of RDD. Therefore, Spark framework is suitable for distributed processing of large-scale data on multiple machines.

Smfi algorithm can be divided into the following four steps:

Step 1: one filtration.

The original data set D is formed into RDD, the map operation forms (item, count) key value pairs for each item in each transaction, the reducebykey operation merges and counts, the

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

filter operation filters the items that meet the support degree to form a frequent item set flist, and the frequent item set is used to prune the original data set to form filterdd1.

Step2: Partition processing.

The FilterRDD1 is projected to be divided into a plurality of partitioned sub-datasets, and a part (suffixtree) is formed in each partition.

Step3: Form candidate frequent itemsets.

In each partition, a deep path Search is performed on the suffixtree, and all the paths obtained form the maximum frequent candidate set (part_MFICS) of the partition. Then, part_MFICS is tested by length first, and the distributed data set part_MFIS is obtained.

Step4: Partition merging.

Combine part_MFIS, and then carry out length-first test to get the final result.

IV. RESEARCH ON DISTRIBUTED DENSITY CLUSTERING ALGORITHM

4.1 Overview of DBSCAN algorithm

A class defined by DBSCAN algorithm is a dense point area separated by sparse point areas, which is a collection of the largest data objects connected in density. The algorithm needs to give two parameters, namely the neighborhood radius threshold Eps and the minimum density threshold Minpts. For each class, each data object must be in the neighborhood radius Eps, and the number of all data objects is the density of the class, and the density must be greater than or equal to the density threshold Minpts.

Relevant definitions are as follows:

Define 1 neighborhood

For an object P in the data set, its neighborhood is the set of all objects in the sphere formed by taking the object P as the center and Eps as the radius, which is written as Formula 1:

$$NEps(m) = \left\{ n \in D \middle| dist(m,n) \le Eps \right\} \quad (1)$$

Define 2 core objects

For any object in the data set, if the number of objects contained in its neighborhood radius is not less than the minimum density threshold, the object is called the core object.

Define 3 boundary objects

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

An object that is not a core object in data set, which is included in Eps range of other core objects, and is generally at the edge of a class, so it is called a boundary object. As shown in fig. 3, a boundary object can be in the Eps range of different core objects. taking minpts = 5 and Eps = r as an example, if there are 8 objects in Eps range of a, then a is the core object, while b has only 3 objects in EPS range, so b is not the core object. But b is contained in the neighborhood of a core object, so b is a boundary object.

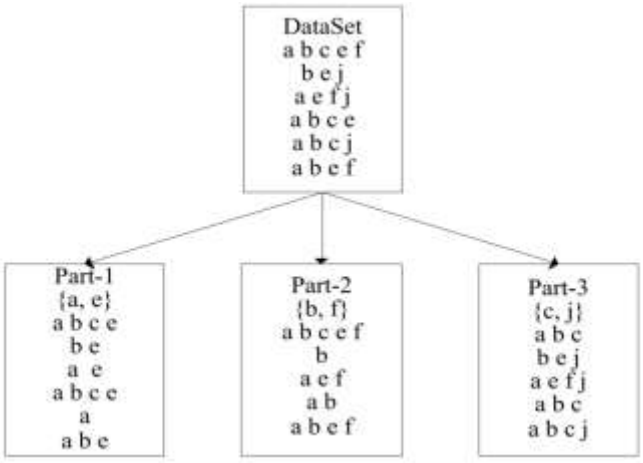

Fig 3: Schematic diagram of core object and boundary object

4.2 Description of distributed SPDBSCAN algorithm based on Spark

Spdbscan algorithm can be divided into the following three steps:

(1) The original data set is transformed into a distributed elastic data set RDD, and the operation of each partition is realized by mapping operator to realize the parallel processing of data partition.

(2) For each partition, the density clustering method of first finding and then merging is used. This method first traverses the data set to form non overlapping groups, and then groups the remaining unclassified points again. These two steps can convert the data set into grouping form, and then search all groups for EPS connected groups, and combine the connected groups to form classes.

(3) According to different situations, this paper proposes an improved merging method, which first merges the groups locally, and finally merges all the partitions globally, so as to improve the efficiency of merging.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

The complexity of the algorithm can be analyzed from the space and time.

(1) From the perspective of space

In the traditional single machine density clustering algorithm, no matter how large the original data set is, it needs to be read into memory at once to cluster globally. If the data set is large, the algorithm can not even run normally. Therefore, in the same time period, the traditional algorithm has high memory requirements for physical hosts, and the size is the size of the whole data set (assuming S). The improved spdbscan algorithm uses the conversion of original data sets from distributed elastic data sets RDD to realize parallel density clustering of different partitions. Therefore, in the same time period, the number of nodes in the distributed cluster is assumed to be N, and for each host node, the memory space occupied by the improved algorithm is only S / N.

From the point of view of the extra storage space occupied by the runtime, the traditional algorithm needs to save all the detected classes and all the objects in them. During the operation of the algorithm, one data object may be contained in multiple classes, and the data objects contained between classes will be redundant. However, the idea of grouping data objects first and then merging them between groups in the improved algorithm is that the data objects contained between groups and the groups contained between classes are non-overlapping and will not occupy extra space. It can be seen that the space complexity of the improved SPDBSCAN algorithm is much smaller than that of the traditional algorithm.

(2) in terms of time

For the traditional DBSCAN clustering algorithm, every time a new core object is added and a class is extended, the whole data set needs to be traversed, and the time complexity is $O(S_2)$ (where S=N*d, d is the data amount in each partition). The larger the data set, the longer the algorithm needs. The improved SPDBSCAN algorithm only processes the data with the size of d=S/N in the same time period, and the time complexity is $O(d^2)$. Moreover, compared with G-DBSCAN algorithm, the calculation method of eps connected groups is improved and the calculation amount is reduced when the clustering results in the partition are merged among groups. When merging classes between partitions, as long as there are eps connected groups between classes between partitions, the remaining groups need not be judged, and the classes are merged directly, which effectively accelerates the processing speed of the algorithm, and the total time complexity is $O(d^2)+O_{mege}$. Therefore, the related strategies and parallelization processing in grouping stage are improved. Compared with G-DBSCAN algorithm, the time complexity is $O(N^2d^2)+ O_{mege}$, so the average time complexity of the improved SPDBSCAN algorithm is smaller than that of G-DBSCAN algorithm.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

V. CONCLUSION

The research of big data analysis is one of the hot issues in the computer field at present. People hope to mine hidden and valuable information through various analysis and processing of big data. However, when dealing with large-scale data, traditional data mining algorithms have some problems, such as long running time, unsatisfactory mining results, and sometimes they can't even run the algorithm. Therefore, it is particularly important to study mining algorithms that can effectively deal with large data sets. In this paper, combined with Spark distributed technology, the algorithm of maximum frequent items of association rules and the density clustering algorithm in clustering analysis are studied, and an improved method which can deal with big data efficiently is proposed. Although this paper has made some achievements in the research of data mining algorithm in distributed environment, there are still some deficiencies and improvements.

ACKNOWLEDGEMENTS

This research was supported by Project of Research and Practice on Higher Education Teaching Reform.(Henan Province2019 ([2019] No.787) :Research and Practice on the Path of Employment and Entrepreneurship Based on Small-scale Customized Productive Teaching Worksite for Prosthetics and Orthopaedics in the Context of Industry-education Integration, Project Number: 2019SJGLX607).

REFERENCES

[1] Qi Juncheng, Liu Bin, Chen Rongchang. Study on X-ray Field Imaging Technology. Acta Physica Sinica, 2019, 68 (02): 96-101
[2] Zhang Sen, Shang Genfeng, Pu Jiexin. Simulation of Accurate Extraction of Underwater Target Image of Marine Resources. Computer Simulation, 2018, 035 (008): 188-193260
[3] [11]Liu ya, AI Haizhou, Xu Guangyou. A moving target detection and tracking algorithm based on background model. Information and Control, 2002, 12: 14-19
[4] [12]Pan Quan, Cheng Yongmei, Du Yajuan. Discrete moment invariant algorithm and its application in target recognition. Acta Sinica Sinica, 2001, 23 (001): 30-36
[5] [13]Ming Dongping, Luo Jiancheng, Shen Zhanfeng, et al. High resolution remote sensing image information extraction and target recognition technology. Surveying and Mapping Science, 2005, 30 (003): 18-20
[6] Long C, Li G, Hongxing Y. 3d Dynamic Object Reconstruction Technology Based on Light Field Rendering. Journal of the University of Chinese Academy of Sciences, 2009, 26 (6): 781-788

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 78-88
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

[7] Yang Fan, Yuan Yan, Zhou Zhiliang. Study on Evaluation Method of Optical Field Camera Imaging Quality. Modern Electronic Technology, 2011, 191(1):146-156

[8] Liu Yanlei, Yuan Libo. Multi Directional Fourier Contour Recognition Method for Steep Edge of Objects. 2013, 7:2(2):729-734.

[9] Tang Yi, Liu Weining, Sun Dihua. Application of Improved Time Series Model in Expressway Short-term Traffic Flow Prediction. Computer Application Research, 2015, 32 (1): 146-149

[10] Wan Ying, Han Yi, Lu Hanqing. Discussion on moving target detection algorithm. Computer Simulation, 2006, 023 (010): 221-226