ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 151

2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)

A Survey of Clustering Algorithms Based on Parallel Mechanism

Xingang Wang

BeiHang University, BHU, No. 37 Xueyuan Road, Haidian District, Beijing, P.R. China, 100083

***Abstract*—With the rapid development of the Internet and mobile Internet, the amount of data and information generated by people has dramatically increased. The demand for rapid processing of data by computers has become increasingly urgent. Clustering analysis is one of the most important data processing methods. Existing clustering algorithms have a high time complexity in calculating the center point and a large amount of resource consumption and poor execution efficiency in the serial processing of mass data. Therefore, efficient and accurate parallel clustering algorithms need to be studied. This paper introduced the parallel mechanism and its computing platform, summarized the existing parallel clustering algorithms, classified them according to clustering methods, and discuss the key technologies and platforms proposed in the existing work.**

Keywords—parallelism mechanism; parallel computing platform; parallel clustering algorithm

I. INTRODUCTION

Due to the rapid development of the Internet and mobile Internet, the amount of data and information generated by people has dramatically increased. Therefore, it is particularly important for computers to have the ability to process data rapidly and obtain useful information. Data mining technology came into being, and it refers to extracting or mining useful knowledge from large-scale data. The continuous development of high-performance computing methods has led to the gradual improvement of information processing systems for massive data, and resource sharing, data storage computing and system coordination work and other issues have been further resolved. Clustering analysis is a common analysis method in data mining. In recent years, although some progress has been made in the research work of clustering analysis in improving the operation efficiency of the algorithms, the algorithms themself have some limitations, that is, the process of analyzing and processing massive data is very complicated. Therefore, the improvement and optimization of the algorithms can not meet the real-time requirements during the using process. In current researches, parallel clustering analysis has drawn wide attention and become a research hot spot. Using multiple computers to process data in parallel greatly improves the operation efficiency of the algorithms. This article aims at the parallel clustering analysis method, and summarizes the existing work and the corresponding core ideas.

Section 2 of this article introduces the parallel mechanism and computing platforms. Section 3 summarizes the existing clustering algorithms based on parallelization. Section 4 summarizes the work of this article.

II. PARALLEL MECHANISM

A. Overview of Parallel Mechanism

Parallel mechanism uses the idea of divide and rule, that is to split and process large-scale problems or tasks[1,2]. A number of divided sub-problems or sub-tasks are assigned to different processors. Through cooperation between processors, parallel processing is performed to reduce the task time and improve the processing efficiency of large-scale problems. In the parallel mechanism, cluster communication is a very important concept. The working way of cluster communication is to connect a group of independent computers through LAN or other ways to work closely together, forming an efficient parallel processing network system. In such a network system, each computer is usually called a node, each node has its own processing system, and their roles in the network system are equal to each other. The communication between nodes uses messaging for data exchange, and parallel programs are used for resource management[3-5].

Parallel mechanism has a very wide range of applications in many fields, such as data mining[17], image processing, game design and so on. In the aspect of data mining, using multi-threads to search and process massive data in parallel can save a lot of running time and greatly improve the processing efficiency. In the aspect of image processing, the parallel processing of graphic information achieves a high speed and meets the real-time requirements at the same time, ensures the image quality. In the aspect of game design, multiple threads are used to parallel perform scene changes, character movement, equipment display tasks, which greatly improves the response speed and optimizes the gaming experience.

B. Research Status of Parallel Mechanism

After decades of development and research, the parallel mechanism is constantly improving and has been widely used in high performance computing. The parallel computer architecture, parallel programming and communication technology are all inseparable in the parallel mechanism. The merits and demerits of the parallel algorithm have a direct impact on the performance of the computer, and the communication technology plays an extremely important role in information transmission. In the mid-1960s, single instruction multiple data computers and single-processor multi-pipelined vector machines appeared one after another. The use of multi-channel programs allows the CPU and I/O operations to switch among multiple programs. After that shared memory based parallel computers occurred, are the same time parallel

processing software and environment were developed, the parallel mechanism has gradually matured and become the mainstream technology. Currently, there are mainly three kinds of parallel systems that are widely used: MPP parallel machine, DSM parallel machine, and workstation cluster. A new generation of parallel systems are also in the process of development. In addition to enhancing the performance of a single computer node, the connection method between nodes has become a research hotspot.

Due to the continuous development of parallel mechanism, the research of parallel clustering algorithms are also becoming deeper and deeper, but at present the main focus is on the calculation and processing of data pairs. Shao Yan put forward a DFA parallel construction algorithm based on multi-thread parallel reading and writing[11], which greatly accelerated the data processing speed, but this technology is not suitable for clustering algorithms. Zhang Xueping et al. put forward the idea of K-medoids parallelization algorithm based on MapReduce programming model[12], which improved the computational efficiency. Experiments show that the K-medoids parallelization algorithm has good clustering results and scalability, but the best effects have not achieved for the initialization process and the allocation process.

C. Parallel Computing Platforms

Currently, some computing platforms designed specifically for massive data processing have been widely used, such as Hadoop, Spark, Storm and so on.

Hadoop[6] is an earlier parallel computing platform that can be used for massive data computing and processing, and the two most important components are MapReduce and HDFS. The programming model of MapReduce includes two functions: Map and Reduce, and specific functionalities are achieved through the design of different functions, without the need for users to be qualified with experiences in parallel programming. The HDFS system is located at the bottom of Hadoop and is mainly used for parallel storage files. It has high fault tolerance, high throughput, and low hardware requirements. Due to the increasing amount of data and frequently updated applications, Hadoop was unable to meet people's needs, at this time theHadoop 2.0 version came into being with the introduction of YARN. The emergence of Resource Manager and Application Master has rationalized the resource management of the entire cluster and significantly improved resource utilization.

Spark is a real-time interactive parallel computing platform that has almost all the advantages of Hadoop, but performs better on some workloads[7]. Due to the use of RDD computing mode and cluster working mode, the cache of the initial data and the intermediate data are easier to control, and the iterative computing ability is greatly improved. Therefore, Spark is very suitable for algorithms of multiple iterations. Spark has become another ecosystem that includes Spark SQL, Spark streaming, Data Mining and MLlib, and Graphx[8]. Real-time, interactive and autonomy design makes the Spark platform widely used, such as Taobao, Jingdong and other famous Internet companies are beginning to use Spark platform for massive data processing.

Storm is a stream processing parallel computing platform based on Clojure programming language. It has the advantages of low latency, real-time performance and multi-language interface support. It is widely used in advertising, real-time analysis, product recommendation, online machine learning and other fields[9,10]. The programming framework is similar to that of MapReduce, and uses the newly created Spout and Bolt for batch processing and parallel processing of stream data. The biggest difference between is that the data processing in Storm is real-time rather than separate batch processing. Storm has many functional components, where, the component Nimbus is responsible for the allocation of tasks and monitoring the status of nodes. The component Supervisor is responsible for monitoring the tasks on the local node and start or shut down the task as needed. The two components are coordinated to ensure the normal operation of the system.

III. PARALLEL CLUSTERING ALGORITHMS

A. Classification Based Clustering Algorithms

Ma Xiaohui et al[13] proposed a parallel K-medoids clustering algorithm. K-medoids clustering algorithm is actually a variant of K-means algorithm[14], the main difference is that the center points are chosen in different ways. The K-medoids algorithm selects the center point whose sum distance to the other data points in the current cluster is the smallest, while the parallel K-medoids algorithm proposed in this article performs initialization by selecting a random center point. The idea of the algorithm is as follows: firstly, k initial data nodes are randomly selected as the cluster centers; secondly, the distances from the remaining nodes to every center points are calculated and each node is assigned to the nearest cluster, during this procedure the Round-Robin strategy is adopted to allocate this task to a number of threads[15,16], and the overall assignment result is obtained by merging the assignment results of each thread; and then loop through the remaining data points in the cluster as the new center point, and compare the communication cost of the new center and the original center; finally determine the data point with the smallest communication cost as the final center point, and return the current cluster and the points contained in the cluster. In the algorithm, a maximum value is used to limit the number of cycles, that is, the number of times for selecting a new center point is controlled no greater than the maximum value.

Since the initial centers of clusters in the parallel K-medoids clustering algorithm in reference [13] are randomly selected, there are some influences on the quality of the clustering results, reference [13] also proposed an improved parallel K-medoids clustering algorithm. The algorithm uses the tag co-occurrence frequency to divide the tags first, and then chooses the clustering centers according to the tag clusters, which improves the algorithm in reference [13].

B. Density Based Clustering Algorithms

Chen Min et al[18] proposed a new parallel clustering algorithm for computer clusters: GP-DBSCAN[19,20]. DBSCAN clustering algorithm is based on density, and it is mainly used for clustering analysis of the databases involving geographic information. Due to the limitation of the application

of DBSCAN algorithm in large-scale database, that is, the memory and the computation speed cannot meet the actual requirements, reference [18] improves the algorithm by dividing the database and using dynamic load balancing. The general steps of the GP-DBSCAN algorithm are as follows: firstly, the values of the parameters EPS and MinPts are determined after the input to the database; secondly, the database is reasonably divided according to the distribution characteristics of the database in each dimension; and then, the partitioned parts are assigned to compute nodes for clustering according to the DBSCAN algorithm; finally, each computing node uploads the result of its own clustering analysis to the master node for integration to obtain the final complete aggregation result. In order to improve the utilization rate of computer cluster resources, reference [18] adopted a dynamic load balancing strategy. Each computing node uploads its own state to the master node periodically, once the upload time exceeds the predetermined period, the master node determines that its state is dead, and assigns the task to other compute nodes.

C. Hierarchy Based Clustering Algorithms

Wang Min et al.[21] improved the traditional CURE algorithm and proposed a parallel CURE algorithm under Binary-Positive. The CURE algorithm is a new hierarchical clustering algorithm, which chooses clustering in the area between the centroid and the surrounding abnormal points[22,23]. Instead of using a single centroid to represent a cluster, it randomly selects multiple hash points and moves them toward the centroid according to the contraction factor, which in turn determines the shape of the cluster. Due to fact that the CURE algorithm samples data points randomly, once the data distribution is uneven, these data points are not representative, and the accuracy of the clustering results will be affected. In view of the above situation, the binary data (Binary-Positive) method was used in reference [21] to select the attribute values of the original data to filter invalid data. That is to say, when a certain attribute of the data exists and its value is reasonable, it is marked as 1, otherwise marked as 0. If the attributes of multiple data exist at the same time but differ greatly, then the average value is taken to adjust the threshold. The processed data is divided into several parts and assigned to multiple Map functions for clustering analysis. Since the order of calculation between Map functions is not related to the calculation of the distance between clusters, the calculation can be executed in parallel. Then the Reduce functions combine the clusters of the same type to get the final clustering results.

D. Grid Based Clustering Algorithms

Zhang et al.[24] studied how the mean approximation method of grid clustering algorithm can be implemented in parallel under the MapReduce framework[25,26]. The average approximation method of grid clustering algorithm integrated both density-based method and grid-based method, and the basic idea is that the width of the grid cells d and the number k of initial clusters are given in advance. First, the raw data is preprocessed, and the data are divided according to the width of the grid cells, the grid cells containing data are marked, and the density and the centroid of the data in the cell are calculated. Then, select the centroid of the k cells with the highest data density as the initial cluster center, assign each of them to the nearest cluster, and then calculate the average value of every centroid in each cluster as the new cluster center. Finally, delete the clusters containing fewer data points than the given parameter, then the the final clustering result is obtained. In order to achieve parallelization, reference [24] adopted the Hadoop MapReduce parallel programming mode, which divides the data set to be clustered into multiple data blocks and distributes them to different hosts for processing, which saves the computation time and improves the efficiency of clustering analysis. Since the data blocks are independent of each other, the parallel clustering results are irrelevant to the calculation order of the data blocks, that is, the final clustering result is consistent with the original result.

E. Model Based Clustering Algorithms

Reference [27] combined the EM algorithm based on Gaussian mixture model with the MapReduce programming framework, and proposed a parallel EM clustering algorithm based on Hadoop platform[28-30]. At present, the initialization of EM algorithm mainly has the following categories: random initialization, K-means initialization and hierarchical clustering initialization. Since the known initialization methods are not very good, reference [5] gives a density-based MergeC initialization method, which selects the best candidate set of the center part of each cluster, and performs weighted combination on them to calculate the initial parameters of the Gaussian mixture model. The main steps of the parallel EM algorithm are as follows: using the data set and the pre-calculated parameter values, firstly execute step E to calculate the probability that each sample data belongs to each cluster, and then execute step M, which puts all the probability values obtained in step E into the corresponding formula to calculate a new round parameters of Gaussian mixture model, so as to achieve the purpose of updating the center and distribution of each cluster. Step E and step M are repeated until convergence is achieved. According to the model parameters achieved at this time, the data is divided into the belonging clusters to obtain the final clustering result. Reference [27] also uses the Hadoop MapReduce framework to implement the parallel mechanism, which designed two MapReduce phases to represent the execution of step E and step M respectively. The experimental results show that the algorithm can achieve clustering analysis of massive data, and the use of parallelization greatly shortens the calculation time and improves the operating efficiency.

IV. CONCLUSION

Based on a brief overview of the paralleliszation mechanism and computing platforms, this article focuses on the parallelization based clustering algorithms, classifies the existing parallel clustering algorithms and analyzes the basic ideas of the representative algorithms in each class. Parallel clustering algorithms significantly improve the data processing speed, shorten the running time and improve the resource utilization rate. Although the parallel mechanism has now been widely used, but it is still under continuous development, and there are still many problems to be further improved. More deeper researches are required on how to optimize parallel computing platforms, better integrate the computing platforms

and clustering analysis, and further improve the data processing speed and operational efficiency.

REFERENCES

[1] Coulouris G, Dollimore J, Kindberg T. Distributed systems : concepts and design - 4th ed.[J]. Programmable Controllers, 2011, 18(95):182–231.

[2] Werner Vogels. Web services are not distributed objects. Internet Computing, IEEE, 2003, 7(6): P59-66

[3] Zhou Xiaofeng, Wang Zhijian. Overview of Distributed Computing Technology [J]. Computer Times, 2004 (12): 3-5.

[4] Malewicz G, Austern M H, Bik A J C, et al. Pregel: a system for large-scale graph processing[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2010:135-146.

[5] Isard M, Budiu M, Yu Y, et al. Dryad: distributed data-parallel programs from sequential building blocks[J]. Acm Sigops Operating Systems Review, 2007, 41(3):59-72.

[6] Xia Jingbo, Wei Zekun, Fu Kai, et al.Research and application of Hadoop technology in cloud computing [J] .Computer Science, 2016, 43 (11): 6-11

[7] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets[C]// Usenix Conference on Hot Topics in Cloud Computing. USENIX Association, 2010:10-10.

[8] HU Jun, HU Xian-de, CHENG Jia-xing.Modeling of Big Data Hybrid Computing Based on Spark [J] .Application of Computers, 2015,24 (4): 214-218.

[9] Github Inc. Storm Wiki [EB/OL]. [2014-11-02]. https://github.com/apache/storm.

[10] Li Chuan, E Haihong, Song Meina.Research and application of real-time computing framework based on Storm [J]. Software, 2014 (10): 16-20.

[11] SHAO Yan. Research on the parallelization of regular expression matching algorithm [D]. Beijing University of Posts and Telecommunications, 2012.

[12] Zhang Xueping.K-Medoids parallel algorithm based on MapReduce.Computer Applications, 33 (4), 2013,1023-1025.

[13] Ma Xiaohui.A Parallel K-medoids Clustering Algorithm.Computer Computer and Applications, 3 (5), 2015,874-876.

[14] Mao Jiali, Wan Min, Chen Hua month. Parallel k-means algorithm in cluster environment [J] .Yibin University, 2007, 12: 91-93.

[15] Goldberger J, Tassa T. A hierarchical clustering algorithm based on the Hungarian method[J]. Pattern Recognition Letter, 29(1). 2008, 1632-1638.

[16] Park H S, Jun C H. A simple and fast algorithm for k-medoids clusting[J]. Expert Systerm with Applications,36(2). 2009, 3336-3341.

[17] Jiang Yuan, Zhang Zhaoyang, Qiu Peiliang et al .. Clustering algorithm for data mining.