Research on Cloud Computing and Its Application in Big Data Processing of Distance Higher Education

http://dx.doi.org/10.3991/ijet.v10i8.5280

Guolei Zhang, Jia Li, and Li Hao
Agricultural University of Hebei, Baoding, Hebei, China

***Abstract*—In the development of information technology the development of scientific theory has brought the progress of science and technology. The progress of science and technology has an impact on the educational field, which changes the way of education. The arrival of the era of big data for the promotion and dissemination of educational resources has played an important role, it makes more and more people benefit. Modern distance education relies on the background of big data and cloud computing, which is composed of a series of tools to support a variety of teaching mode. Clustering algorithm can provide an effective evaluation method for students' personality characteristics and learning status in distance education. However, the traditional K-means clustering algorithm has the characteristics of randomness, uncertainty, high time complexity, and it does not meet the requirements of large data processing. In this paper, we study the parallel K-means clustering algorithm based on cloud computing platform Hadoop, and give the design and strategy of the algorithm. Then, we carry out experiments on several different sizes of data sets, and compare the performance of the proposed method with the general clustering method. Experimental results show that the proposed algorithm which is accelerated has good speed up and low cost. It is suitable for the analysis and mining of large data in the distance higher education.**

***Index Terms*—Cloud computing, Hadoop, Map-Reduce, Distance Higher Education, Parallel k-means clustering algorithm**

I. INTRODUCTION

In the development of information technology the development of scientific theory has brought the progress of science and technology. The progress of science and technology has an impact on the educational field, which changes the way of education. The arrival of the era of big data for the promotion and dissemination of educational resources has played an important role, it makes more and more people benefit. Under the background of cloud computing, big data of education has ushered in the climax of development and expansion. The efficiency and reliability of cloud computing make the mining and analysis of big data more efficient in the educational field. With the data processing efficiency significantly improved, the big data play an increasingly important role in the distance education and teaching practice.

The construction of the modern distance education resources includes the media material library, the question bank, the case base, the courseware storehouse ,the network curriculum, the teaching system which is suitable for many kinds of teaching mode, and the development of the modern distance education management system. Modern distance education teaching system is composed of a series of teaching tools to support a variety of teaching mode, including learning system, teaching system, teaching resources editing system, counseling and answering system, examination system, evaluation system, communication tools, virtual experiment system and search engine, etc.. These teaching tools are based on the distance education resource pool. The resource pool includes the information center of the distance education resources, the resources of different subjects, different regions and different schools. Teaching resource editing system is the basis of distance education. At present, the overall level of distance education system in China is relatively low. In order to have a fully functional resource pool, we must find a way to efficiently deal with the big data of educational resources [1-4].

The research of cluster analysis has a long history. For decades, its importance and the use of the cross of other research directions have been affirmed. Clustering is one of the important research contents in data mining, pattern recognition and so on. It has an extremely important role in the identification of the intrinsic structure of the data. Clustering in pattern recognition is applied to speech recognition, character recognition, etc.. Clustering algorithm in machine learning is applied to image segmentation and machine vision. It is also used for data compression and information retrieval in image processing. Another major application of clustering is data mining (multi relational data mining), Spatio-Temporal database applications (GIS, etc.), and the heterogeneous data analysis[5-7]. In order to realize the clustering analysis of big data, some scholars had realized the parallel clustering algorithm based on the distributed programming model. The Map Reduce parallel design of K-means algorithm was realized by Jiang Xiaoping and Zhao Weizhong[8-9]. Lu Weiming et al. proposed a distributed affinity propagation clustering algorithm based on Map Reduce [10]. A distributed network data clustering algorithm based on Map Reduce was presented in Chen Dongming's article [11]; Qi Weizhong et al. proposed a distributed spectral clustering algorithm based on Map Reduce [12]. A massive text clustering method based on cloud computing was proposed by Cao Zewen et al. [13]. A graph clustering algorithm based on Map Reduce was proposed in the literature [14]. A clustering algorithm of Map Reduce parallelization based on grid was proposed by [15].

Because the traditional K-means clustering algorithm has the characteristics of randomness, uncertainty, high time complexity, and it does not meet the requirements of large data processing. In this paper, a parallel K-means clustering algorithm based on cloud computing platform Hadoop is studied, and the method and strategy of the algorithm are given. This paper is organized as follows: in section 2, we introduce the basic knowledge of Hadoop architecture. Introduce the basic theory of clustering method in the 3 section. In the fourth section, we construct a parallel K-means clustering method. In the fifth section, we first carry out experiments on several datasets of different sizes. Then, we compare the performance of the clustering method which is accelerated with the common clustering method. Finally, the analysis of the experimental results is given. The last two sections are the conclusion and the references.

II. HADOOP

Hadoop is a framework for distributed computing, which is organized by the Apache foundation. It uses low equipment to build large computing pool, in order to improve the speed and efficiency of big data analysis. Hadoop imitate and implement the Google cloud computing technology, which includes HDFS (Hadoop distributed file system), Map-Reduce, HBase, and ZooKeeper,etc..

(1)Hadoop Common: It supports the public part of the Hadoop. The common Hadoop is the lowest level of the module; its main function is to provide a variety of tools for other subsystems.

(2)HDFS: It is a master/slave structure, which consists of a Name-Node and several Data-Nodes. Name-Node is responsible for the management of the file system metadata, while the Data-Node is responsible for storing the actual data. Fig. 1 shows the architecture of Hadoop-based distributed systems. As shown in the figure, they generally consist of a master and workers. The master is in charge of maintaining workers, building query execution plans for user queries, and orchestrating the entire query execution process. Workers are responsible for processing user queries according to the query plans.

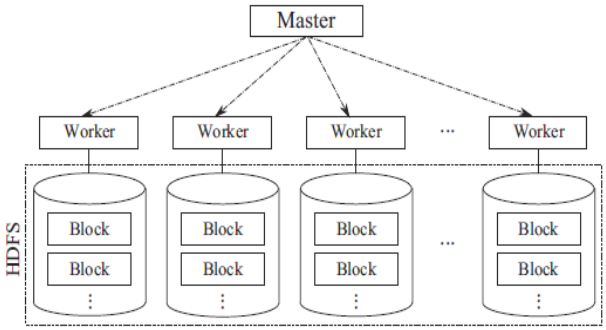

Figure 1. Architecture of Hadoop-based distributed systems

(3) Map-Reduce: Its main function is to decompose the task, and to summarize the results of a sub task. Among them, the Map decomposed the main task into multiple sub tasks, and the Reduce put the results of multiple sub task processing. Then, we can get the final results. Map-Reduce decomposes big data into many small data sets. The big data is the railway passenger flow data which is read from the HDFS. Each small data set is processed in parallel, and then stored in a distributed database.

(4) Hive: It is a data warehouse tool that provides the query function of the SQL statement.

(5) Hbase: It is a distributed database based on column storage model.

III. THE BASIC KNOWLEDGE OF CLUSTERING ALGORITHM

A. Clustering criterion

Assume that the n samples $x_j \in R^N \ (j = 1,2,\mathrm{L}\ ,n)$ is divided into c classes.

The $i = 1,2,\mathrm{L}\ ,c$ and $j = 1,2,\mathrm{L}\ ,n$ is defined as:

$$\mu_{ij} = \begin{cases} 1 & if\ the\ ith\ sample\ belongs\ to\ the\ class\ j \\ 0 & Otherwise \end{cases}$$

Then the matrix $\mu = (\mu_{ij})$ has the following properties:

Set

$$\mu \in \{0,1\}, \sum_{i=1}^{c} \mu_{ij} = 1, \ j = 1,2,\mathrm{L}\ ,n$$

Let n_i denote the number of samples contained in the i th class,

Then $n_i = \sum_{j=1}^{n} \mu_{ij}, \ i = 1,2,\mathrm{L}\ ,c$

Let $\overline{x}_i \in R^N$ denote the center of the i th class

Then $\overline{x}_i = \frac{\sum_{j=1}^{n} \mu_{ij} x_j}{\sum_{j=1}^{n} \mu_{ij}} = \frac{1}{n} \sum_{j=1}^{n} \mu_{ij} x_j, \ i = 1,2,\mathrm{L}\ ,c$

Therefore, the within-group variation of the i th class is as follows:

$$S^i(\mu) = \sum_{j=1}^{n} \mu_{ij} \|x_i - \overline{x}_i\|^2$$

Then the whole within-group variation is

$$S(\mu) = \sum_{i=1}^{c} S^i(\mu) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij} \|x_j - \overline{x}_i\|^2$$

The above is the classical criterion formula of within-group sum of squared error (WGSS). The k-means clustering algorithm is aimed at finding $\mu^* = (\mu_{ij}^*)$ that makes $S(\mu)$ has the minimum value.

B. The type of clustering

According to the clustering criterion of data in clustering, there are many kinds of clustering algorithms. Clustering algorithm has a variety of classification methods; they can be roughly divided into 4 categories. They are hierarchical clustering algorithm, the partition clustering algorithm, density and grid clustering algorithm, and other clustering algorithms. As shown in Figure 2.

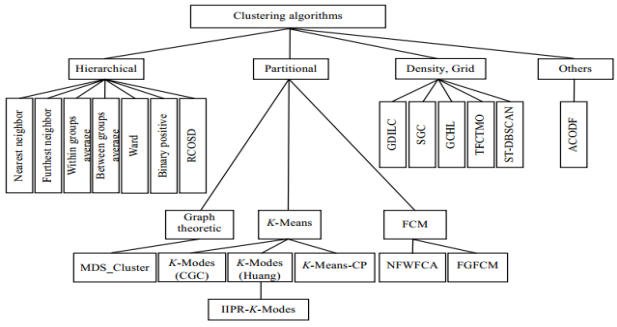

Figure 2. The classification chart of clustering algorithms

IV. DESIGN OF PARALLEL K-MEANS CLUSTERING ALGORITHM BASED ON HADOOP

A. Serial K-mean algorithm

Design of parallel algorithm is based on the Hadoop design, the user's the main work is design and implementation of map and reduce functions, including the type of key value of input and output, and the map and reduce functions specific logic.

The steps of the serial K-mean algorithm are:

Step1: The k samples are selected as the initial center point of the cluster;

Step2: Iteration. Firstly, according to the center point coordinates of each cluster, each sample is assigned to the nearest cluster. Secondly, the center point of the cluster is updated, that is to calculate the mean value of all the samples in each cluster;

Step3: until convergence.

As can be seen from the K-means algorithm, the main calculation work in the algorithm is to assign each sample to the nearest cluster, and the operation of different samples is independent of each other. Therefore, we consider the implementation of this step in parallel. In each iteration, the algorithm performs the same operation. The parallel K-means algorithm can perform the same operations on the Map and Reduce in each iteration. The k samples are randomly selected as the central point, and they are stored in a file on the HDFS as a global variable. Next each iteration is composed of 3 parts: Map function, Combine function and Reduce function.

B. Parallel K-mean algorithm

A massively parallel and scalable implementation of a parallel k-means algorithm was developed for cluster analysis of very large datasets (Figure 3). The algorithm works exactly the same way as the standard serial k-means algorithm. However, every process operates only on its chunk of the dataset, carries out the distance calculation of points from the centroids and assigns it to the closest centroid. Each process also calculates the partial sum along each dimension of the points in each cluster for its chunk for the centroid calculation. At the end of the iteration, a global reduction operation is carried out, after each process obtains the information, to calculate the new cluster centroids. Iterations are carried out until convergence, after which the cluster assignments are written to an output file.

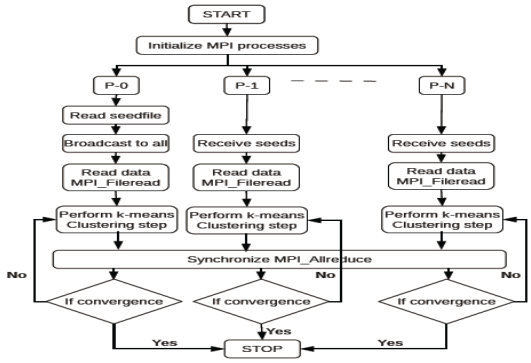

Figure 3. Flowchart for parallel k-means algorithm

V. EXPERIMENT AND RESULT ANALYSIS

According to the students' normal and final exam results, the data mining technology based on clustering can find many internal factors of distance education, such as the students' learning attitude, learning habits, weak links, and family work and so on. The results of the analysis are applied to guide students to study and daily teaching. In this way, not only saves much time to practice, but also can get good learning effect. In the scores analysis module of the distance education system, the function of the intelligent tutoring system based on clustering is added, which can analyze the students' test scores. The results of the analysis include the selection of the courses, the recommended books, and the characteristics of the categories. According to the results of the analysis, students can also carry out targeted learning. So it can improve the intelligence of the whole system. In this paper, we carry out experiments on the big data processing capability of the distance education system. Performance of the parallel k-means algorithm and implementation was evaluated using three example datasets. These datasets (Table 1) have been used in this study to test a broad combination of small to large n and k clustering problems and to test the computational efficiency and scalability of the implementation.

Work reported in this study was carried out using Jaguar which is a Cray XT5 supercomputer. It consists of 10,000 compute nodes of dual hex-core AMD Opteron processors running at 2.6 GHz, 16 GB of DDR2-800 memory, and a Sea Star 2+ router. It contains a total of 150,000 processing cores, and 50 TB of memory. Next, we give three comparison charts of the computational performance in different data sets, which are the accelerated parallel clustering algorithms and the traditional clustering algorithms.

Figure 4-6 shows the scaling of the three datasets of increasing size for increasing numbers of clusters, k=10, 25, 50, 100, 500, and1000. Good performance was achieved by the algorithm in terms of simulation time for all the datasets. The educational resource dataset contains the largest number of data; it is ten times bigger than the information system datasets. As we can see, the processing time of the parallel clustering algorithm which is accelerated is less than that of the traditional clustering method in the three kinds of data sets. It can be seen that, with the increase of the nodes of clusters, the performance will be significantly improved.

TABLE I.
THE DATASETS

Dataset	Dimensions	Records	Dataset Size
Information system datasets	15	50000000	10G
Students basic information datasets	20	20000000	40G
Educational resource datasets	25	100000000	100G

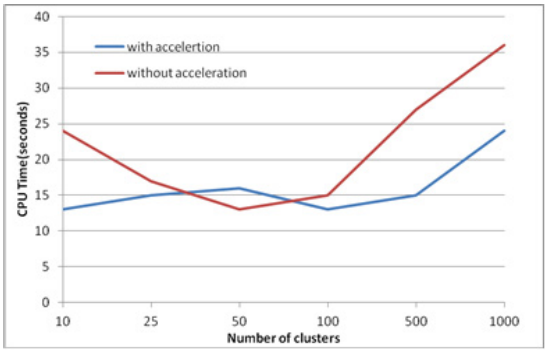

Figure 4. Information system datasets: with and without acceleration

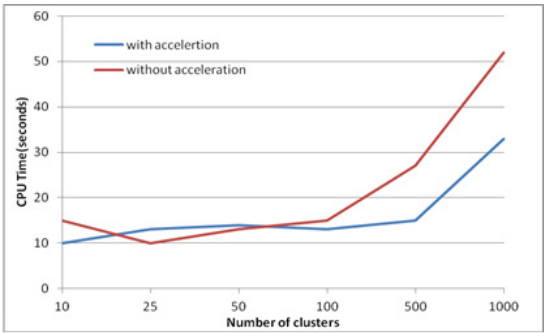

Figure 5. Students basic information datasets: with and without acceleration

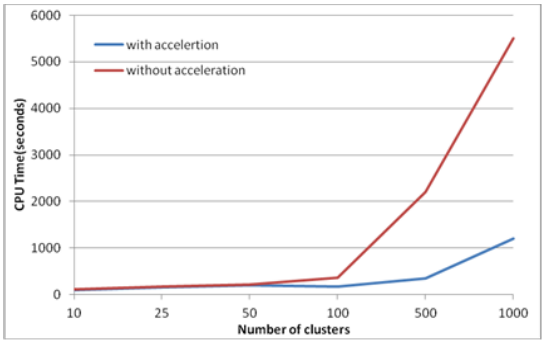

Figure 6. Educational resource datasets: with and without acceleration

VI. CONCLUSION

Clustering algorithm can provide an effective evaluation method for students' personality characteristics and learning status in distance education. However, the traditional K-means clustering algorithm has the characteristics of randomness, uncertainty, high time complexity, and it does not meet the requirements of large data processing. In this paper, we study the parallel K-means clustering algorithm based on cloud computing platform Hadoop, and give the design and strategy of the algorithm. Then, we carry out experiments on several different sizes of data sets, and compare the performance of the proposed method with the general clustering method. Experimental results show that the proposed algorithm which is accelerated has good speed up and low cost. It is suitable for the analysis and mining of large data in the distance higher education.

REFERENCES

[1] The Big Data[OL]. http://baike.baidu.com/view/6954399.htm.

[2] West, Darrell M. Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. Governance Studies at Brookings [R].Washington: Brookings Institution, 2012:1-10.

[3] Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics [OL]. Http://www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf.

[4] Anthony G. Picciano. The Evolution of Big Data and Learning Analytics in American Higher Education [J].Journal of Asynchronous Learning Networks, 2012,3 (3):9-20.

[5] Jain AK, Flynn PJ. Image segmentation using clustering. In: Ahuja N, Bowyer K, eds. Advances in Image Understanding: A Festchrift for Azriel Rosenfeld. Piscataway: IEEE Press, 1996. 65−83.

[6] Cades I, Smyth P, Mannila H. Probabilistic modeling of transactional data with applications to profiling, visualization and prediction, sigmod. In: Proc. of the 7th ACM SIGKDD. San Francisco: ACM Press, 2001. 37−46. http://www.sigkdd.org/kdd2001/

[7] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999, 31(3): 264−323. http://dx.doi.org/10.1145/331499.331504

[8] Jiang Xiaoping, Li Chenghua, Xiang Wen et al. Parallel implementing K-MEANS clustering algorithm using MapReduce programming mode[J]. J.Huazhong Univ. of .Sci & Tech.(Natural Edition),2011,39(1):120-124.

[9] Zhao Weizhong, Ma Huifang, Fu Yanxiang et al. Research on parallel k-means algorithm design based on Hadoop platform [J]. Computer Science, 2011, 38(10):166-168.

[10] Lu Weiming, Du Chenyang, Wei Baogang. et al. Distributed affinity propagation clustering algorithm based on MapReduce [J]. Journal of Computer Research and Development, 2012, 49(8): 164-174.

[11] Chen Dongming,Liu Jian, Wang Dongqi et al. Distributed Clustering Algorithm for Network DataBased on MapReduce [J]. Computer Engineering, 2013, 39(7):76-82.

[12] Zhong Q,Lin Y, Zou J, et al.Parallel spectral clustering based on MapReduce[J]. ZTE Communications,2013,2(11):30-37.

[13] Cao Zewen, Zhou Yao. Design and Implementation of JP Algorithm Based on MapReduce [J]. Computer Engineering, 2012, 38(24):14-16.

[14] He Guoxiong. Research and Implementation of MapReduce-based Graph Clustering Algorithm [D] . Chang Sha: Hunan University, 2012.

[15] Zhang Lei, Zhang Gongrang, Zhang Jinguang. MapReduce Parallelization Research of a Clustering Algorithm Based on Grid [J]. Computer Technology and Development, 2013, 23(2):60-64.

AUTHORS

Guolei Zhang, male, was born in Feb.1978. He obtained a master degree of agronomy from Agricultural University of Hebei in July 2007, and now works at College of Information Science and Technology in Agricultural University of Hebei, No.289, Lingyusi Street, Baoding, Hebei, China as the vice dean and vice professor, studying in application of novel science and theories in management of college students. (123866119@qq.com).

Jia Li, female, bore in January, 1984, Agricultural University of Hebei. She obtained a Master degree in Agricultural Extension in July, 2015. At present, she works as instructor at College of Modern Science and Technology of Agricultural University of Hebei, No.2596, Lekai of south Street, Baoding, Hebei, China. Her title is lecturer. Her research field is education and management of college students. (280937678@qq.com).

Li Hao, female, was born in Apr.1983. She obtained master degree of engineering from Agricultural University of Hebei, and now works at Agricultural University of Hebei, Personnel Division, No.289, Lingyusi Street, Baoding, Hebei, China as a lecturer, studying in ideological and political education (holy0228@126.com)

This paper was supported by Innovation Research Center for Practical Education of Socialist Core Values of Agricultural University of Hebei, Key Research Base of Humanities and Social Sciences of Institutions of Higher Learning in Hebei Province and by people's livelihood research project of Hebei Province (Grant No.201501311). Submitted 07 November 2015. Published as resubmitted by the authors 05 December 2015.