

Design and implementation of a parallel algorithm based on Hadoop platform

Qingnian Zhang*, Zhao Chen, Zihui Wang

Wuhan University of Technology, China

Received 1 October 2014, www.cmnt.lv

Abstract

Existing clustering algorithm is transplanted into the Hadoop cloud computing platform, through the low price on the computer cluster nodes dynamically allocate huge amounts of data distributed task, solve the enterprise needs a large amount of data storage and the problem of real time analysis results. Graphs programming model can help developers to quickly realize the parallel clustering, and do not need too much to understand the specific underlying communication realization. This article will improve the clustering algorithm, which is transplanted into graphs on the programming model, realize the parallel design, and through the error sum of squares criteria such as function test and verify the reliability of the parallel algorithm. Under the Hadoop cluster composed of four machines respectively samples of different sizes of data clustering analysis, proves that the parallel algorithm of Hadoop platform on the large data applications better speedup and scalability.

Keywords: hadoop platform, mapreduce, clustering, k-mean

1 Introduction

Hadoop open source distributed cloud computing platform by the Hadoop distributed file system and the graphs of programming model. Existing clustering algorithm is transplanted into the Hadoop cloud computing platform, through the low price on the computer cluster nodes dynamically allocate huge amounts of data distributed task, solve the enterprise needs a large amount of data storage and the problem of real time analysis results [1, 2]. Graphs programming model can help developers to quickly realize the parallel clustering, and do not need too much to understand the specific underlying communication realization [3, 4].

Through learning clustering algorithm in data mining, this paper implemented the Hadoop distributed platform parallel Kmeans algorithm, effectively solves the massive user data classification problem.

2 Data clustering application framework design

This system adopts the Hadoop platform Hadoop distributed file system to store user data, with the clustering algorithm into graphs programming for data clustering. Figure 1 is the flow chart for the application of this system framework, which mainly consists of data collection module, data preprocessing module, data storage module and data mining analysis module [5, 6].

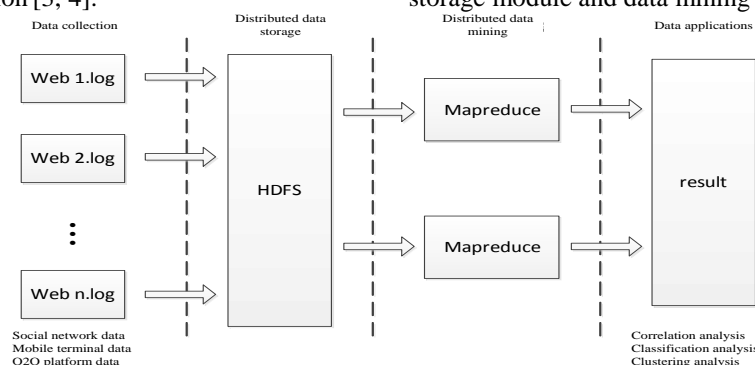


FIGURE 1 System frame

Mapreduce programming model can effectively solve the parallel processing data in the distributed storage and fault tolerance, how could the Hadoop platform to realize the big data mining work, the key is the transfer of the

mining algorithm to graphs programming model. Under the framework of the k-means algorithm is transplanted into graphs, the concrete implementation process programming model as shown in Figure 2.

*Corresponding author e-mail: jieyang509@163.com

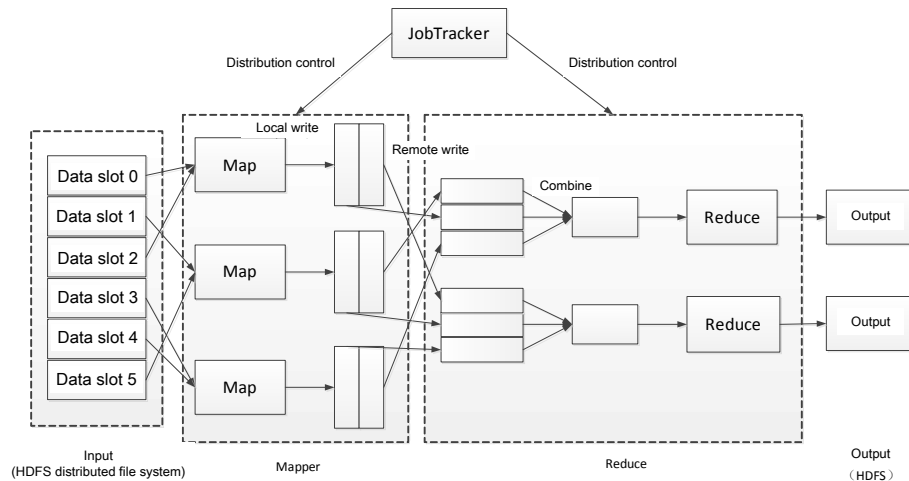


FIGURE 2 Mapreduce programming model specific implementation process

3 The site data clustering application framework

3.1 HARDWARE AND SOFTWARE CONFIGURATION

Website user behaviour data not only are there more big quantity, type, and for real-time analysis of the results, according to the characteristics of this system USES Hadoop platform Hadoop distributed file system to store user data, with transplanted into graphs programming under the clustering algorithm to cluster the user consumption behaviour [7, 8]. By the results of clustering, the difference of customers to provide high quality service, in view of the current marketing activities of the problem, provide favourable data support.

Using 4 PC computer in ubuntu12.04 platform structures, a small cluster with four nodes. Choose one of

the highest frequencies of CPU a master, as the main nodes at the same time set up the master and the other three as a slave node. Master node configuration is 1 GB memory size, from the node configuration memory size is 512 m, all 20 GB hard drive size distribution. Every node in the cluster need to install the same version of Hadoop and under the same path JDK version, through a router connection to the machine, every day in the same local area network (LAN). Cluster structures, different from the pseudo distributed mainly in two aspects: the static IP Settings and SSH communications. After launch the Hadoop in a small cluster environment, on the master there should be the JobTracker, SecondaryNameNode, the NameNode, DataNode and TaskTracker five daemons, while slave1, slave2 and slave3 should have DataNode and TaskTracker two daemons. Table 1 below is the Hadoop cluster software and hardware configuration.

TABLE 1 hardware and software configuration information

	master	slave1	slave2	slave3
cpu	2.5GHz	1.79GHz	1.83GHz	1.93GHz
memory	1G	512M	512M	512M
hard disk	20G	20G	20G	20G
operating system	Ubuntu12.04, Windows XP			
open tools	Hadoop-1.0.2, Eclipse, JDK1.6			

3.2 THE HADOOP NODE CONFIGURATION

Hadoop is mainly composed of two basic modules: graphs programming model and Hadoop distributed file system. Graphs programming model is mainly used in distributed applications, this model is a JobTracker and several TaskTracker. After the client send a job application to the JobTracker, JobTracker is responsible for the operations of initialization, job scheduling and monitoring TaskTracker task. As long as find task failure, JobTracker will restart it, guarantee the stability of the cluster. TaskTracker is running on the slave service, responsible for executing the JobTracker distribution of parallel tasks [9, 10]. HDFS distributed file system is used for distributed storage, as the underlying support of graphs, the nodes in the cluster consists of one NameNode and more DataNode. The NameNode's role is to manage the file system namespace,

the function of the DataNode is according to the need to store and retrieve data block. This experiment is configured with one NameNode and four DataNode, Table 2 for the specific planning of Hadoop environment.

TABLE 2 graphs and HDFS node configuration information

	master	JobTracker/TaskTracker
Mapreduce	slave1	TaskTracker
	slave2	TaskTracker
	slave3	TaskTracker
	master	NameNode/DataNode
HDFS	slave1	DataNode
	slave2	DataNode
	slave3	DataNode

3.3 SIMILARITY MEASURE METHOD

This paper uses the modified method of similarity measure in the map function to achieve the effect of normalized processing. Traditional k-means algorithm is used in the graphs, the map function defined in the two samples x_a , x_b similarity is the use of Euclidean distance, the smaller

$$s(x_a, x_b) = \sqrt{\sum_{j=1}^n (x_{aj} - x_{bj})^2}, \text{ smaller } s = (x_a, x_b)$$

shows that the greater the similarity between samples. Sample calculation and $s = (x_a, x_b)$ size of each centre, and clustering to the smallest clusters of $s = (x_a, x_b)$ distance class. In order to achieve the result of the normalized, modify the judgment criterion of the map function, to redefine the sample similarity distance is:

$$s'(x_a, x_b) = \sum_{j=1}^n \frac{1}{1 + |x_{aj} - x_{bj}|}. \quad (1)$$

Similarity between each attribute was defined as the reciprocal of the Manhattan distance plus 1, the greater the value of $s = (x_a, x_b)$ shows the greater the similarity between the sample data. By improving the similarity measure of $s = (x_a, x_b)$ calculated data sample and each centre distance, the result value, the greater the explain data samples is more similar with the centre. Due to the improved similarity measure $s = (x_a, x_b)$ to map samples of each attribute to between 0 and 1, guarantee the contribution of each attribute function in the same frame of reference, have the effect of the normalized. Such clustering result also has more explanatory.

4 Parallel algorithm design

4.1 CONSTRUCTION OF THE CLUSTER

Cluster is the master node and slave nodes distribution on different computers. In the practical application of Hadoop cluster scale is compared commonly big, but the main purpose of this paper is to validate the Hadoop in the usability of the website user data clustering, this simulation environment for the four nodes Hadoop cluster, transplanted clustering algorithm into Hadoop platform.

4.2 DESIGN AND IMPLEMENTATION OF PARALLEL ALGORITHM

In the Hadoop platform the key step for realizing the parallel Kmeans algorithm is to design the Map function and a Reduce function. Every process of Mapreduce is equivalent to a serial clustering algorithm in an iterative process. The Map function is to get the data object distance to the centre node, and distributing the data object to the

nearest cluster class, and marking the category ID. The Reduce function completes new centre in the same cluster class value calculation. Relatively new clustering centre, compared with the previous clustering centre, when the error is less than the convergence value, the clustering process is over, otherwise the circulation process of Mapreduce. For the convenience of the recording and processing the data and Mapreduce calculation model uses the line form for storage, as there is no correlation, piece of data to be processed data line can be shard. To ensure the independence of the clustering process each iteration the data can be distributed to store the data node.

4.2.1 The map function design.

The main content of the Map function is calculated each data to all the selected centre distance, and put each data tag to the centre of the shortest distance, belong to the same data centre is a bunch of classes. The Map function is < key, value > way to input and output. The input data is the centre of the previous iteration (or the centre of the random selection) and all the sample data. Input function < key, value > object relative to the corresponding is the key data in the data file offset of the starting point and the value corresponding to the data object coordinate values of each dimension. Output data is each data object belongs to the ID, the output function < key, value > value or represent data objects in the coordinates of each dimension, and the key value indicates that the clustering centre of the data object. The Map function to realize pseudo code:

```
void map(LongWritable key, Text value){
    mis_distance=getEuclideanDistance(point,cluster[0]);
    for(int i=0; i<k; i++){
        If(getEuclideanDistance(point, cluster[i])<mis_distance){
            mis_distance=getEuclideanDistance(point, cluster[i]);
            currentCluster_ID=i;
        }
    }
    }intermediate_output(currentCluster_ID, point);
}
```

4.2.2 Reduce function design

The main content of the Reduce function is to the same cluster the data averaged class, get a new centre, if the centre of two adjacent results don't happen deviation, clustering results are produced. If the result deviation, this time the Reduce output centre as the initial centre of the next iteration. Reduce function of the input is the output of the Map function, the results of the same key value is assigned to the same Reduce calculation (i.e., the same cluster class assigned to a Reduce in). Corresponding form input function < key, value > for < cluster ID of the list (belong to the centre of the data objects) >, calculate the same in each Reduce the number of the class and each component and cluster computing the mean of each component of the clustering centre as a new file. Output function < key, value > for < > cluster ID, sample vector. The Reduce function implementation of pseudo code:

```

void reduce(Writable key, Iterator<PointWritable> points){
long num=0;
Float sum=0.0f;
While(points.hasNext()){
PointWritable current_point=points.next();
num+=current_point.getNum();
for(int i=0; i<dimension; i++)
sum[i]+=current_point[i];
for(int i=0; i<dimension; i++)
mean[i]=sum[i]/num;
}result_output(key,mean);
}

```

5 Similarity measure experiment and analysis

5.1 CLUSTERING CRITERION

Clustering analysis of the target is to belong to the similar research object into a class, belong to the same object of a class as similar as possible, and not the same kind object is different, as far as possible, the basic clustering criterion contains three functions.

5.1.1 The error sum of squares criterion function.

Data sets $R = \{r_1, r_2, r_3, \dots, r_n\}$, by clustering into k clusters, namely $R_1, R_2, R_3, \dots, R_k$, each class contains data $m_1, m_2, m_3, \dots, m_k$ for a number of clusters, defined error sum of squares criterion function is:

$$j_a = \sum_{j=1}^k \sum_{n=1}^{m_j} \|r_n^j - c_j\|^2, \quad (2)$$

r_n^j of them belong to the same cluster class j of sample data. While c_j said j -th clusters mean all the samples in the class, the class cluster centre, its representation is:

$$c_j = \frac{1}{m_j} \sum_{j=1}^{m_j} r_j, j = 1, 2, 3, \dots, k. \quad (3)$$

Error variance criterion function mainly calculation belong to the same cluster samples of a class and the class of cluster to Euclidean distance to the centre of the square, when the smaller the value of j_a , suggests that cluster-heads data between difference is smaller, the clustering effect is good. In the process of clustering iterations, similar data will be allocated according to the judgment of the distance to the same cluster, j_a function present a descending trend. However, only use this a judgment standard is not enough, the error sum of squares criterion suitable for sample data density is bigger, small differences between the clusters of judgment, in the case of within the cluster sample difference is very big, the error sum of squares criteria often cannot render good effect [11].

5.1.2 The weighted average of the sum of squares of the criterion function

This function is shown by the following:

$$j_b = \sum_{j=1}^k p_j f_j^*. \quad (4)$$

The p_j is the prior probability, of the total number of samples for m , assigned to the j -th class cluster of number for m_j , is the representation of a prior probability method is:

$$p_j = \frac{m_j}{m}, j = 1, 2, 3, \dots, k \quad (5)$$

and f_j^* said j -th cluster-heads mean value of the square of the distance between two objects, because m_j number within the cluster is, the combination of two number is $m_j(m_j - 1)$, expression is:

$$f_j^* = \frac{2}{m_j(m_j - 1)} \sum_{r \in R_j} \sum_{r' \in R_j} \|r - r'\|^2. \quad (6)$$

In view of the error variance criterion function cannot very good judgment sample data of markedly different defects, weighted average of the sum of squares of good make up for it. Rule of the main computing the sum of squared distance between classes in the same cluster sample. As the clustering iterations, similar samples gathered in the same cluster, cluster sample within the distance between the smaller and smaller, the weighted average of the sum of squares of the criterion function said local area density, will present a downward trend.

5.1.3 The distance between the class and the criterion function

The above two criteria are judgment within the cluster, the similarity between data, and the distance between the class and the criterion function is more focused on different types of separation degree, the rule of the well of the distance between each cluster are described, the expression is as follows:

$$j_c = \sum_{j=1}^k (c_j - c_{all})^T (c_j - c_{all}). \quad (7)$$

Expression of c_j said all the samples in the class is the first j cluster averages, c_{all} said average of all the sample data. In the process of clustering iterations, each cluster degree of dissimilarity between more and more obvious, the distance between the class criterion function present a rising trend.

Taken together with the increase of clustering iterations, the error sum of squares criterion function and

weighted average of the sum of squares of the criterion function as class data cluster distance more and more close, property is more and more similar, function values can present a downward trend, show the same kind object difference as small as possible. And the distance between the class criterion function with data between clusters of separation, function values can present a rising trend, show differences between different objects as large as possible.

5.2 THE RESULT OF THE EXPERIMENT

This experiment from UCI Machine Learning Repository site selection of the iris data set, wine data set, the data set and the vehicle data sets upload the four sample data to hadoop distributed file system. Then the graphs programming under the framework of Kmeans algorithm to cluster analysis, and use and square error criterion function, weighted average of the sum of squares of the criterion function. The distance between the class and criterion function on the process of clustering analysis, observation is related to the changes in the process of

iteration function, verify the reliability of the parallel clustering algorithm.

The iris sample set a total of 150 data, each data contains 4 kind of attribute, the whole data set can be divided into 3 clusters; Wine sample set a total of 178 data, each data contains 13 kinds of attributes, the whole data set can be divided into 3 clusters; Seed data samples with 210, each sample has eight kinds of attributes, samples can be classified into 3 clusters; Vehicle data samples with 94, each sample has 18 kinds of attributes, samples can be classified into four clusters. Because the Kmeans algorithm is randomly selected from the initial value, different initial value selection may get different clustering results, so this paper 20 times for each criterion function clustering operation, a randomly selected as the experimental results. At the beginning of the iterative function change is more obvious, therefore this thesis excerpts from the first five iterations as trend analysis. Table 3-6 of iris respectively, wine, seed, vehicle data aggregation class effect.

TABLE 3 Iris data aggregation class effect

Number of iterations	Error sum of squares	Weighted average sum	Distance between the class sum
1	3.143×10^3	1.278	10.354
2	2.653×10^3	1.169	10.803
3	2.487×10^3	1.147	11.028
4	2.067×10^3	1.137	11.286
5	2.048×10^3	1.120	11.652

TABLE 4 Wine data aggregation class effect

Number of iterations	Error sum of squares	Weighted average sum	Distance between the class sum
1	8.580×10^7	3.155×10^4	2.068×10^5
2	6.815×10^7	2.821×10^4	2.832×10^5
3	7.233×10^7	2.821×10^4	2.847×10^5
4	7.045×10^7	2.773×10^4	2.823×10^5
5	6.814×10^7	2.729×10^4	2.854×10^5

TABLE 5 Seed data aggregation class effect

Number of iterations	Error sum of squares	Weighted average sum	Distance between the class sum
1	6.045×10^4	9.971	30.261
2	4.672×10^4	8.874	30.921
3	3.438×10^4	7.162	31.992
4	2.467×10^4	6.448	31.968
5	2.243×10^4	5.973	31.594

TABLE 6 Vehicle data aggregation class effect

Number of iterations	Error sum of squares	Weighted average sum	Distance between the class sum
1	1.051×10^7	2.047×10^4	1.019×10^5
2	6.991×10^6	1.324×10^4	1.140×10^5
3	7.115×10^6	1.184×10^4	1.309×10^5
4	6.475×10^6	1.137×10^4	1.440×10^5
5	5.798×10^6	1.132×10^4	1.468×10^5

By the form can be found on the four sample data shows that with the increase of the number of iterations.


Clustering similar sample data set in the same class makes the error sum of squares criterion function. The weighted

average of the sum of squares of the function presented a decreasing trend, and differences between class and class increases gradually, the distance between the classes and functions presented an increasing trend. But the three functions is not monotone changing, in the process of clustering iterative search criterion function of the uncertainty of the optimum path, the criterion function results may fluctuate, but the overall trend in line with the desired results. At the same time can also be observed in

three previous iterations, the change of the function value is relatively obvious, to late function value change is small, the clustering results. The experiment proves that the graphs under the programming model of parallel Kmean algorithm, has good clustering effect. Under the Hadoop cluster composed of four machines respectively samples of different sizes of data clustering analysis, proves that the parallel algorithm of Hadoop platform on the large data applications better speedup and scalability.

References

- [1] Lu H, Hu T-t 2012 Research on Hadoop Cloud Computing Model and its Applications *Networking and Distributed Computing(ICNDC)* 59-63
- [2] Yu Hong, Wang D 2012 Mass Log Data Processing and Ming Based on Hadoop and Cloud Computing *Computer Science & Education (ICCSE)* 197-202
- [3] Singh B, Singh H K 2010 Web Data Mining Research: a Survey *Computational Intelligence and Computing Research (ICCIC)* 1-10
- [4] Kala Karun A, Chitharanjan. K A 2013 Review on hadoop -HDFS infrastructure extensions *IEEE Conference on Information and Communication Technologies* 132-7
- [5] Dean J, Ghemawat S 2004 MapReduce: symplified data processing on large clusters *OSDI* 1-12
- [6] Zhou J, Liu Z 2008 Distributed clustering based on k-means and CPGA *Fuzzy Systems and Knowledge Discovery* 2 444-7
- [7] Hai M, Zhang S, Zhu L, Wang Y 2012 A survey of distributed clustering algorithms *Industrial Control and Electronics Engineering* 1142-5
- [8] Hirzel M, Andrade H, Gedik B 2013 IBM Streams Processing Language: Analyzing Big Data in motion *International Business Machines Corporation* 1-11
- [9] Garlasu D, Sandulescu V, Halcu I, Neculoiu G A 2013 Big data implementation based on Grid computing *Roedunet International Conference (RoEduNet)* 17-9
- [10] Kala Karun. A, Chitharanjan K A 2013 Review on hadoop -HDFS infrastructure extensions *IEEE Conference on Information and Communication Technologies* 132-7
- [11] Pakhira M K 2009 Clustering large databases in distributed environment *IEEE International Advance Computing Conference* 351-8

Authors	
	<p>Qing Nian Zhang, China</p> <p>Current position, grades: professor of School of Transportation at Wuhan University of Technology since 2002. University studies: PhD degree in Machinery design and theories from the Wuhan University of Technology, China, in 2002. Scientific interest: traffic and transportation planning, optimization and decision making of transportation system, transportation safety management.</p>
	<p>Zhao Chen, born in 1969, Shaanxi Province, China</p> <p>Current position, grades: PhD degree in Logistics management at Wuhan University of Technology, China. University studies: MS degree in Transportation management engineering from Wuhan University of Technology, China, in 2002. Scientific interest: parallel computing on hadoop platform, optimization decision.</p>
	<p>Zihui Wang, born in 1990, Wuhan, China</p> <p>Current position, grades: M.S. degree in Electronics and Communication Engineering at Wuhan University of Technology, Wuhan, China. University studies: B.S. degree in Electronic Communication Engineering from Wuhan University of Technology, Wuhan, China, in 2012. Scientific interest: signal processing on Hadoop platform, pattern recognition.</p>