The Analysis and Implementation of the K - Means Algorithm Based on Hadoop Platform

Liu Xiang wei[1]

[1] PLA University of Foreign Languages, Luoyang Henan, China

Correspondence: Liu Xiang Wei, PLA University of Foreign Languages, Luoyang Henan, 471003 China. E-mail: liuxwletter@163.com

Received: November 3, 2016 Accepted: November 23, 2016 Online Published: January 29, 2018

doi:10.5539/cis.v11n1p98 URL: http://dx.doi.org/10.5539/cis.v11n1p98

Abstract

In today's society has entered the era of big data, data of the diversity and the amount of data increases to the data storage and processing brought great challenges, Hadoop HDFS and MapReduce better solves the these two problems. Classical K-means algorithm is the most widely used one based on the partition of the clustering algorithm. At the completion of the cluster configuration based on, the k-means algorithm in cluster mode of operation principle and in the cluster mode realized kmeans algorithm, and the experimental results are research and analysis, summarized the k-means algorithm is run on the Hadoop platform's strengths and limitations.

Keywords: K-means, clustering, Hadoop

1. Introduction

With the coming of information age, more diverse data sources, data volume is growing rendering geometry type. A relational database system has been unable to such a large data storage management, previous processing system is not able to deal effectively with these data. Therefore, how to mass data storage and processing has become a time problem, in this respect has created such as Hadoop, Spark big data technology.

K - Means algorithm as a kind of clustering algorithm based on partition and application is very extensive, operation time, quick speed, easy to explain and has better clustering effect. K - means algorithm the basic idea is: centering on the k point in space clustering, classification for the most close to their objects. Through iterative method, the value of the successive update the clustering center, until the best clustering results are obtained. For these technologies in the research, design and implementation, to grasp the big data analysis and processing, has the vital significance.

Hadoop is mainly introduced in this paper the construction of the distributed processing environment, and based on different data to analyze the efficiency of running in a distributed environment, implemented within the framework of graphs Kmeans algorithm, and through the experimental results are analyzed, and improvement opinions are put forward.

2 Hadoop Cluster Environment Configuration

2.1 The Computer Environment Configuration

The Cluster is composed of five computer, 1 master4 slave. In order to facilitate the configuration file copy, all the user name of the machine are hadoop, password for hadoop. Host name is different, the master of M, slave respectively S1, S2, S3, S4. The IP address of the host for continuous distribution

2.2 Hadoop Environment Configuration

2.2.1 Configure the Network

The network configuration steps are as follows:

(1) The first input cd

(2) Enter sudo vim/etc/network/interfaces

(3) Interface documentation to join at the end of the following statements

```
auto eth0
```

```
    iface eth0 inet static
    address 192.168.2.210
netmask 255.255.255.0
```

2.2.2 Configure SSH

Configure SSH mainly include public and private keys:

(1) In four machine input respectively SSH - the keygen -t rsa to generate a digital signature.

(2) Copy the public key to the server host.

(3) Copy from the machine's public key to the host. After SSH directory, need to create authorized_keys file on a host computer to store the public key.

(4) Then enter sudo reboot to restart the computer, make effective configuration file.

(5) After the next to be distributed to each host password-less login within a cluster can be realized.

2.2.3 Install the JDK and HADOOP

Unpack the JDK and HADOOP compression bag, Modify the configuration startup files, After inventory, reboot the system. Boot into the system, check whether successful installation:

Java version - version information will be shown,The hadoop command prompt information will be shown.

2.2.4 Configure Hadoop cluster host M the XML file

Host M is needed in the operation

(1) First enter the hadoop installation directory, enter the mkdir TMP TMP establish the temporary file directory.

(2) Modify the hadoop/etc/hadoop hadoop - env. Under sh, mapred - env. Sh and yarn - env. Sh.

(3) Modify the hadoop/etc/hadoop core - site. Under the XML.

(4) Modify the hadoop/etc/hadoop HDFS - site. Under the XML.

(5) Modify the hadoop/etc/hadoop mapred - site. Under the XML

At this point has been configured on the host M hadoop related documents, and the configuration file on the host S1 and S6 for distribution

2.2.5 The Cluster Start

(1) If you want to start the HDFS file system requires the host M command line, enter the start - dfs.sh

(2) And then type the command on the host M the JPS command can see the namenode and secondarynamenode start, launched a datanode on other slave hosts.

2.3 The client host configuration

(1) Installing a Ubuntu desktop host, configured network and follow these steps to install hadoop and jdk.

(2) The hadoop - eclipse plugin - 2.7.1. Copy the jar to eclipse plugins folder under the root directory

(3) Open the eclipse, select OpenPersepctive Window menu options in the other, select Map/Reduce again.

(4) In the Map/Reduce the Location of a new Loaction, input information of this cluster. After can see DFS file system.

(5) Immediately after the new program and select run on Hadoop.

3. K - Means Algorithm on Hadoop Platform Implementation

3.1 The Feasibility Analysis of K-Means Algorithm

3.1.1 K - Means Algorithm

K - Means algorithm is a kind of typical clustering method based on classification. This method first requires the user to input the number k of clustering and specify the initial clustering center, k begin after clustering. Select one of the first data point, after calculating the data point to the center of the k cluster distance (Euclidean distance, for example), then compared the data points and the size of the k cluster center distance and will point to the smallest distance clustering. For after the completion of all the data points, to calculate the new clustering center (the clustering can be all the points in the average). After defining new clustering center, a new data points of clustering algorithms, until the center of mass is constant or changed little.

3.1.2 Mapreduce Architecture

Graphs as an important part of the hadoop platform, its basic characteristic is divided into two parts, one is a Map, the main role is to Map key/value pair, to cut word frequency statistics for example is the first word, and then give each word value is 1, after the Reduce is a specification of each node Map for results, which is the same key to delete a while, add 1 to the corresponding value, thus eventually get word frequency statistics results. The architecture of the most powerful and most prominent advantage is the work can be similar and not influence each other to break up, then assigned to different node for processing, eventually integrating criterion to get the final result.

Through the calculated data points assigned to each node, and then reduction processing. Under the condition of the cluster scale allows, in a short period of time can be a lot of iterative computation. So graphs architecture is very suitable for K - Means algorithm was implemented. K - Means algorithm is also facing a problem, in a large amount of data, dimension more cases, single machine is difficult to complete relevant operation. So the MapReduce architecture also solves the key problem of K-Means algorithm surviving in the current large data environment.

3.2 Algorithm Implementation Assistance

Algorithm of the auxiliary work to be done mainly by the Helper classes, including the following functions。The self-test section contains the following classes, as shown in Table 1.

Table 1. Self-Test section table

The function name	Parameter	Effect
getOldCenters	inputPath	The old centroid file is read in and stored in ArryList
deleteLastResult	path	Delete the data on the HDFS file system corresponding to path
copyOriginalCenters	src dst	From the local path src, copy the user-defined initial cluster center data to the specified path dst in the HDFS file system
isFinished	oldPath newPath Kpath dtBegIdxPath threshold	First, the old cluster center is read through the oldPath. Since this function is executed after an iteration of the K-Means algorithm, the newPath corresponds to the current clustering result, and then the current clustering result is new. , And then compare the distance between the new clustering center and the old clustering center is less than the set threshold threshold. If it is less than, it returns true, which means the cluster center does not need to be updated, and has reached the precision. If it is greater than, update the cluster center and return false to prepare for the next iteration.

3.3 Run Document

Kmeans internal consists of the following categories:

(1) KmeansMapper；

(2) KmeansReducer；

(3) runKmeans

(4) main；

Among them, some functions such as shown in table 2.

Table 2. Some functions table

The function name	Parameter	Effect	Special instructions
map	Key values context	In acquiring oldCenters file storing old clustering center, through the point of each need clustering and clustering center distance calculations, decided to the cluster to which cluster center, and record.	The main is to call the Mapper class, convenient for K - means algorithm

reduce	Key Values context	After judgment, if need to Reduce the operation, then call this function to Reduce the results of the Map operation, main is to merge the same key value, synchronous update data of all nodes	The main Reducer is call class, convenient for K - means operation
runKmeans	Args isReduce	First to determine whether a number of parameters meet the conditions, if does not meet the error; If meet, Map operation, after the judgment is the need to Reduce the operation, need to Reduce the operation, do not need to, is set up after the input and output path to the output of the results.	
main	args	This function to delete the last run results first, then copy the user defined initial clustering center to the program. Into the circulation, after the end of the loop condition is reached the maximum number of iterations or new clustering center and clustering center of the old distance less than a specified threshold. Loop body perform content is, the first to print the current execution of the number of iterations, is to run after K - Means algorithm. Whether the end is through isFinished function return values.	

On the cluster, the program flow chart in Figure 1 as follows.

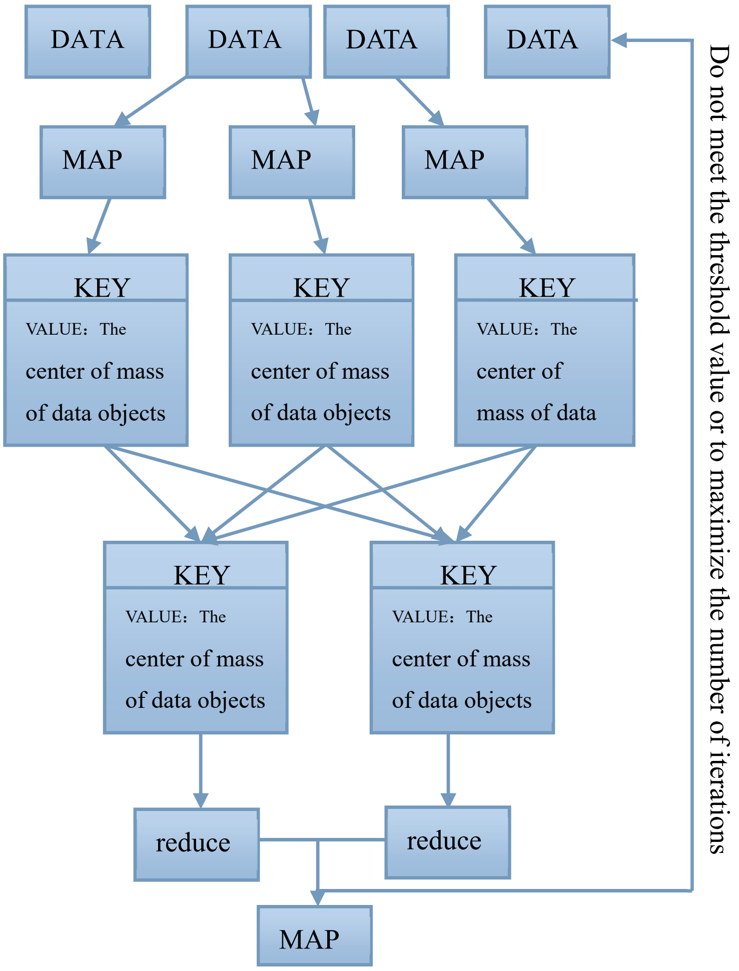

DATA
DATA
DATA
DATA
MAP
MAP
MAP
KEY
VALUE：The center of mass of data objects
KEY
VALUE：The center of mass of data objects
KEY
VALUE：The center of mass of data
KEY
VALUE：The center of mass of data objects
KEY
VALUE：The center of mass of data objects
reduce
reduce
MAP
Do not meet the threshold value or to maximize the number of iterations

Figure 1. The program flow

4. Experimental Process Analysis

4.1 Experimental Data

The data were randomized to a total of 179 groups of 14 randomized data.

4.2 Experimental Procedure

In the process of K-means algorithm experiment, by adjusting the difference of clustering centers, the threshold is adjusted to achieve the clustering purpose.

1. In the above data for K-Means algorithm, the results shown in Table 3:

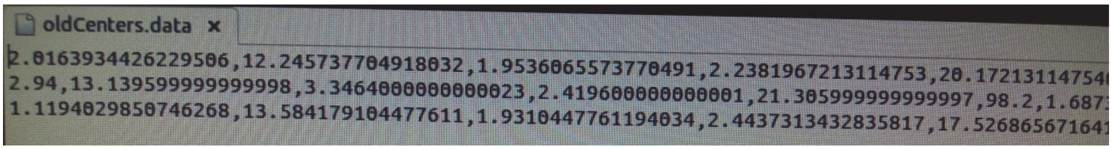
Table 3. The number of clustering centers, the threshold, and the running time of the algorithm

Number of cluster centers	Threshold=0.1	Threshold=0.5	Threshold=1
K=3	94626ms	98234ms	101134ms
K=4	97573ms	103986ms	102949ms

It can be seen from Table 1 that the larger the threshold, the larger the number of clustering centers and the longer the running time of the algorithm.

2. After Kmeans clustering, we finally calculated the data contained in each cluster:

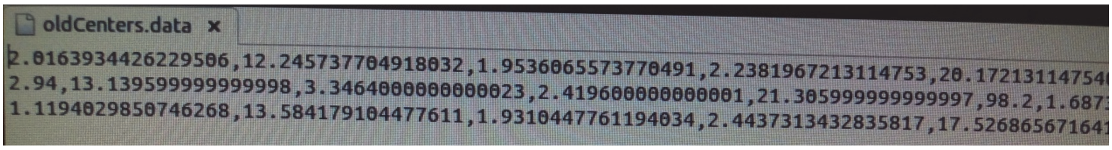
(1) When the number of cluster centers is 3, the threshold is 0.1, and the number of iterations is limited to 20, Fig. 2 shows the three centroids of clustering results.



```
2.0163934426229506,12.245737704918032,1.9536065573770491,2.2381967213114753,20.17213114754
2.94,13.139599999999998,3.3464000000000023,2.4196000000000001,21.305999999999997,98.2,1.687
1.1194029850746268,13.584179104477611,1.9310447761194034,2.4437313432835817,17.52686567164
```

Figure 2. Clustering results

(2) When the number of cluster centers is 4, the threshold is 0.1, and the number of iteration cycles is limited to 20, as shown in Fig. 3



```
2.0163934426229506,12.245737704918032,1.9536065573770491,2.2381967213114753,20.17213114754
2.94,13.139599999999998,3.3464000000000023,2.4196000000000001,21.305999999999997,98.2,1.687
1.1194029850746268,13.584179104477611,1.9310447761194034,2.4437313432835817,17.52686567164
```

Figure 3. Clustering results

The results of the operation, that is, the data centers of the three clusters are shown in Table 4 and Table 5.

Table 4. Operation result

K=3	Cluster center 1	Cluster center 2	Cluster center 3
Threshold=0.1	67	61	50
Threshold=0.5	67	61	50
Threshold=1	67	61	50

Table 5. Operation result

K=4	Cluster center 1	Cluster center 2	Cluster center 3	Cluster center 4
Threshold=0.1	36	68	59	15
Threshold=0.5	36	68	59	15
Threshold=1	36	68	59	15

4.3 Analysis of Results

The results show that the data size of each cluster center is almost the same when the thresholds are not very large, and the accuracy of the data classification is also relatively stable when the number of clusters is small. And the classification results are consistent. After observing the data of 20 clustering centers after 20 iterations, it is found that the smaller thresholds do not affect the clustering center results when the initial centers are consistent. This shows that the stability of K-Means algorithm is very high, itself is also very reliable.

5. Conclusion

In this paper, based on MapReduce distributed computing framework, the K-Means algorithm. The algorithm divides the data first, then distributes it to each slave through the Master. After each slave, the data is processed. Then, the clustering center is clustered synchronously. When the difference between the new cluster center and the old cluster center is less than the threshold or after iterations reach a certain number of times, the operation is stopped and the result of the last clustering is used as the result. Through the comparison and analysis of the results, we can see that the K-Means algorithm based on the MapReduce distributed processing framework has more advantages in dealing with large-scale data.

References

Anchalia, P. P., Koundinya, A. K., & Srinath, N. K. (2013). MapReuce Desigh of K-means Clustering Algorithm. *International Conference on Information Science & Applications, 2013,* 1-5.

Li, L. Y., Dong Y. M., Kong, Y. Zh., & Qiuli, K. (2016) - means algorithm of graphs parallel research. *Journal of Harbin University of Science and Technology, 2016*(1).

Ma, Handa, Hao, X. Y., & Ma, R. M. (2015). Hadoop based parallel PSO-k means algorithm for Web log mining. *Journal of Computer Applications, 42*(z1).

Yuan, X. Y. (2016). ABC_Kmeans clustering algorithm of graphs parallel study. *Computer Measurement and Control, 2016*(1), 252-254.

Zhou, T., Zh., J. Y. L., & Cheng, K. (2013). Means clustering algorithm based on Hadoop is the implementation. *Computer Technology and Development, 23*(7), 18-21.

Copyrights