# K-means Clustering Optimization Algorithm Based on MapReduce

Zhihua Li[1,a], Xudong Song[2,b],WenhuiZhu[3,c], YanxiaChen[4,d *]

*[1]College of Network Engineering, Shijiazhuang Institute of Technology, Shijiazhuang, 050228, China*

*[2]Software Institute, Dalian Jiaotong University, Dalian, Liaoning, 116028, China*

*[3]Beijing Datang Telecom convergence communications Technology Co., Ltd., Beijing, 100029, China*

*[4]Department of Medical Imaging, Dalian Medical University, Dalian, Liaoning, 116044, China*

*[a]email: intand@163.com, [b]email: xudongsong@126.com,[c]email: 553955162@qq.com,[d]email: cyx_dl@126.com, * Corresponding author*

## Abstract

Aiming at the defects of traditional K-means clustering algorithm for big data, this paper provides K-means clustering mining optimization algorithm based on big data, shows a MapReduce software architecture which is suitable for large data processing mechanism, provides an improved method for selecting initial clustering centers and puts forward a K-means algorithm optimization based on MapReduce model. The improved algorithm is applied to the coal quality analysis, the result shows that compared with traditional algorithms, the optimization algorithm improves the efficiency of the algorithm obviously, and the accuracy is also enhanced.

*Keywords:Data Mining, K-means Clustering algorithm,MapReduce, Hadoop*

## 1 Introduction

K-means clustering algorithm is a classical clustering algorithm based on splitting method.Because the theory of the algorithm is reliable, simple and convergent rapidly, K-meansalgorithm is widely used [1][2][3][4][5].

However, with the development of the information society, the data size the data mining tasks faced is more and more big.Even thoughthe traditional clustering mining optimization algorithm have good accuracyin the face of massive data,its time complexity of serial calculation method are high. More how to store, handle these massive amounts of data, and dig outfurther useful knowledge can guide the application become a thorny issue. Aiming at the defects of traditional algorithm, this article proposes an improved method for selecting initial clustering centers and puts forward a K-means algorithm optimization based on Hadoop cloud computing platform. The improved algorithm is applied to the coal quality analysis, the results show that compared with traditional algorithm, the optimization algorithm improves the efficiency of the algorithm obviously, and the accuracy is also enhanced.

## 2 Operation mechanism of MapReduce

Hadoop is an open source distributed computing platform, which mainly consists of distributed computing framework--MapReduce and distributed file systems--HDFS. MapReduce is one of the core components of Hadoop, and it is easy to realize distributed computer programming by MapReduce on Hadoop platform.

MapReduce is a software framework for parallel computing programming model of large-scale data sets, having obvious advantages in dealing withthe huge amount of data.

Operation mechanism of MapReduce is as follows:

(1)Input: MapReduce framework based on Hadoop requires a pair ofMap and Reduce functionsimplementing the appropriate interface or abstract class,and should also be specified the input and output location and other operating parameters.In this stage, the large datain theinput directory will be divided into several independent data blocks for the Map function of parallel processing [6][7].

(2)MapReduce framework puts the application of the input as a set of key-value pairs <key,value>. In the Map stage, the framework will call the user-defined Map function to process each key-value pairs <key,value>, while generating a new batch ofmiddle key-value pairs<key,value>.

(3)Shuffle：In order to ensure that the input of Reduceoutputted by Map have been sorted, in the Shuffle stage, the framework uses HTTP to get associated key-value pairs <key,value> Map outputs for each Reduce; MapReduce frameworkgroups the input of the Reduce phase according to the key value.

(4)Reduce：This phase will traverse the intermediate data for each unique key, and execute user-defined Reduce function. The input parameter is < key, {a list of values} >, the output is the new key-value pairs< key, value >.

(5)Output：This stage will write the results of the Reduce to thespecified output directory location.

Operation mechanism ofMapReduce is shown in Figure 1.

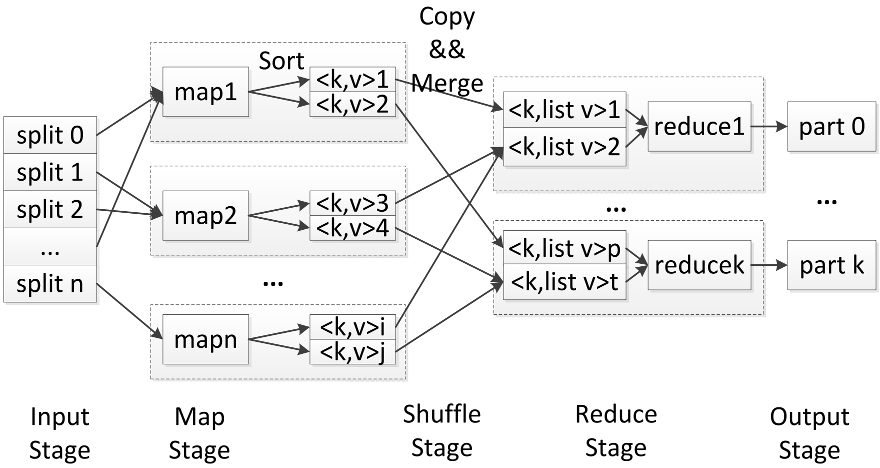


Fig.1.Operation mechanism of MapReduce

## 3 Clustering mining optimization algorithm based on MapReduce

The data set processed by MapReduce should have such characteristics: It can be broken down into many small data sets, and each small data set can be completely parallel processed [8][9][10].The process of K-means algorithm based on Hadoop mainly has two parts, the first part is to initial clustering centers, and divide the sample data set into a certain size of data blocks for parallel processing. The second part is to start the Map and Reduce tasks for parallel processing of algorithm in time, until process gets the clustering results. Its algorithm process is shown in Figure 2.

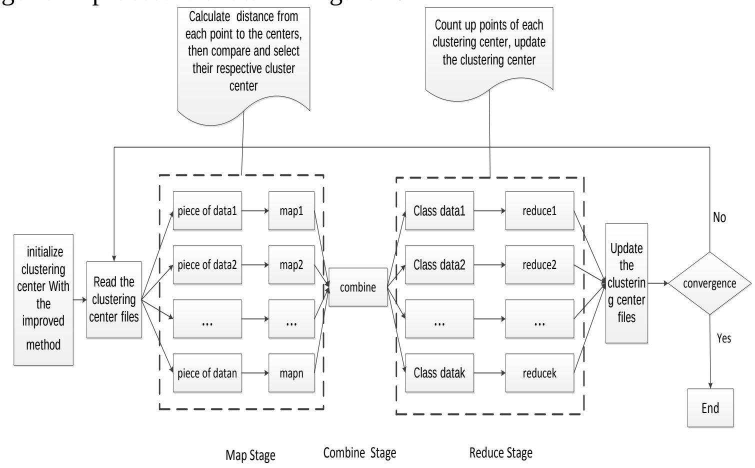

Calculate distance from each point to the centers, then compare and select their respective cluster center
Count up points of each clustering center, update the clustering center
initialize clustering center With the improved method
Read the clustering center files
piece of data1
piece of data2
...
piece of datan
map1
map2
...
mapn
combine
Class data1
Class data2
...
Class datak
reduce1
reduce2
...
reducek
Update the clustering center files
convergence
No
Yes
End
Map Stage
Combine Stage
Reduce Stage

Fig.2. Process of parallel K-means algorithm based on Hadoop platform

The initial clustering centers of traditional algorithm selected randomly, will cause the instability of clustering results. This paper adopts a method of the initial clustering center selection to improve the stability of the results. Optimized K-means clustering algorithm firstly choose k samples to initialize clustering centers according to certain algorithmic rules, then k clustering centers are stored in a file on the HDFS as a global variable [11].

Let cluster sample data set: D= $\{d_i\,|d_i$ ∈R,i=1,2,3,…n}, k cluster centers are showed by $c_1$ , $c_2$ ,$c_3$…$c_k$. Specifically definitions are as follows:

(1)In the data set,distance between any two n-dimensional vector is expressed using Euclidean distance: dist($\boldsymbol{d}_i$,$\boldsymbol{d}_j$)=$\sqrt{(d_{i1}-d_{j1})^2+(d_{i2}-d_{j2})^2+\cdots+(d_{in}-d_{jn})^2}$

(2)Data center of sample points O( $\boldsymbol{d}_i,\boldsymbol{d}_j$ ): O ( $\boldsymbol{d}_i,\boldsymbol{d}_j$ ) = $(\frac{d_{i1}+d_{j1}}{2},\frac{d_{i2}+d_{j2}}{2},\ldots.\frac{d_{in}+d_{jn}}{2})$

(3)The average distance between sample points： averg= $\frac{\sum\mathrm{dist}(\boldsymbol{d}_i,\boldsymbol{d}_j)}{C_n^2}$ (i, j =1,2,3,…n),namely the sum of all the distance between the two sample points divided by the combinatorial number of n sample points.

The initial clustering centers processes are as follows:

(1)Calculate the distance between sample points and store the data in the matrix D

(2)Initialize the set A and the cluster center set C, the minimum distance of sample points is put into the set A, and its center $O_1$is the first initial cluster center in the set C.

(3)Calculate the second nearest point center,then get the distance between this center and $O_1$, compared with averg;If it is less than averg,add the center to set A,and calculate the thirdnearest point center, repeat steps 3; If it is greater than averg, add the center to set C.

(4)Until the number of set C is k.

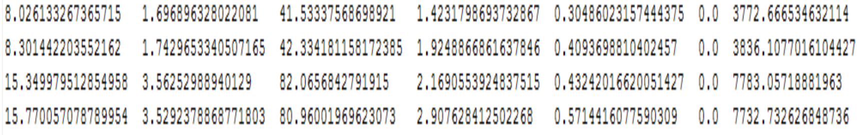
## 4 Experiment and Result Analysis

Ourexperimental data set isfrom a coal group enterprise.This experimentanalyzes results respectively from the effectiveness of the algorithm and speed ratio,using K-means optimization algorithm based on Hadoop for completing the clustering analysis of characteristic data of the coal group enterprise.

We use amachineas the NameNode and JobTracter node; five other machines are DataNode and TaskTracker node. Each node hardware configuration is as follows:

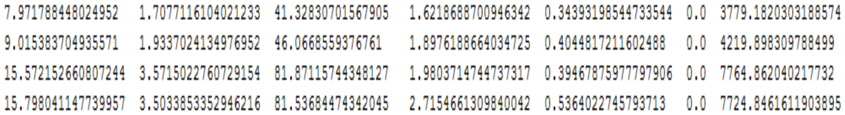
CPU is i5 M 480 @ 2.67 GHz dual-core, memory is 1 g. Hard disk is 250 g / 7200 RPM.

There are 18,038 coal experimental data sample points, using the traditional K-means algorithm and optimization algorithms to test to generate four clusters. Traditional clustering algorithms due to the dependence of the initial cluster centers lead toinstability of clustering results, which clustering results thatdifferent experiments producedare constantly changing, andresults of optimization algorithms remain unchanged. This paper selected two traditional clustering algorithms of results and the optimization clustering results shown in Figure 3, Figure 4, and Figure 5.



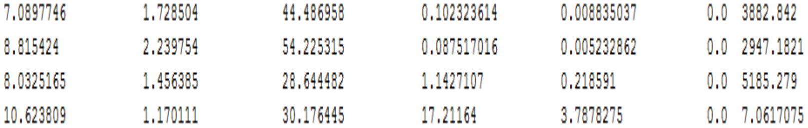
```
8.026133267365715   1.696896328022081   41.53337568698921   1.4231798693732867  0.30486023157444375  0.0  3772.666534632114
8.301442203552162   1.7429653340507165  42.334181158172385  1.9248866861637846  0.4093698810402457   0.0  3836.107701610442
15.349979512854958  3.56252988940129    82.0656842791915    2.1690553924837515  0.43242016620051427  0.0  7783.05718881963
15.770057078789954  3.5292378868771803  80.96001969623073   2.907628412502268   0.5714416077590309   0.0  7732.732626848736
```

Fig.3. Results oftraditional clustering algorithm 1.



```
7.971788448024952   1.7077116104021233  41.32830701567905   1.6218688700946342  0.34393198544733544  0.0  3779.1820303188574
9.015383704935571   1.9337024134976952  46.0668559376761    1.8976188664034725  0.4044817211602488   0.0  4219.898309788499
15.572152660807244  3.5715022760729154  81.87115744348127   1.9803714744737317  0.39467875977797906  0.0  7764.862040217732
15.798041147739957  3.5033853352946216  81.53684474342045   2.7154661309840042  0.5364022745793713   0.0  7724.8461611903895
```

Fig.4. Results oftraditional clustering algorithm 2.



```
7.0897746           1.728504            44.486958           0.102323614         0.008835037          0.0  3882.842
8.815424            2.239754            54.225315           0.087517016         0.005232862          0.0  2947.1821
8.0325165           1.456385            28.644482           1.1427107           0.218591             0.0  5185.279
10.623809           1.170111            30.176445           17.21164            3.7878275            0.0  7.0617075
```

Fig.5. Results ofoptimization clustering algorithm.

Compared with the traditional algorithm,wecan see that the result of optimization algorithm has higher accuracy and stability.

## 5 Conclusion

Aiming at the defects of traditionalK-meansclustering algorithm for big data, this article improved the selection of the initial clustering center firstly, secondlyarticle realized the parallelization of K-means algorithm usingoperation mechanism of MapReduce.Experiments show that the improved algorithm has better effectiveness and higher computational efficiencycompared with the traditional algorithm and the greater the amount of data the more obvious advantages.

## Acknowledgment

This work has been supported by China Scholarship Counciland National Natural Science Foundation of China (No. 61074029).

## References

[1] Su Jin-qi, Xue Hui-feng, Zhan Hai-liang. K-means Initial Clustering Center Optimal Algorithm Based on Partitioning[J]. Microelectronics&Computer, 2009, 26(1):8-11.

[2] Tong Xue-jiao, Meng Fan-rong, Wang Zhi-xiao.Optimization to K-means initial cluster centers [J]. Computer Engineering and Design, 2011, 32(8):2721-2723.

[3] Li Zhengbing, Luo Bin, Zhai Sulan, Tu Zhengzheng.K-means Algorithm

Based on Partition of correlational graph [J]. Computer Engineering and Applications,2013, 49(21): 141-144.

[4] Deng Hai, Tan Hua, Sun Xin. A K-Means Clustering Algorithm of Meliorated Initial Center [J]. ComputerTechnology and Development,2013, 23(11): 42-45.

[5] Zhou Weiben, Shi Yuexiang. Optimization algorithm of K-means clustering center of selection based on density [J]. Application Research of Computers.2012,29(5): 1726-1728.

[6] Zhao Wei-zhong, Ma Hui-fang, Fu Yan-xiang, Shi Zhong-zhi. Research on Parallel K-means Algorithm Design Based on Hadoop Platform [J]. Computer Science.2011, 38(10):166-168.

[7]Ralf Lammel. Google'sMapReduceProgrammingModel-Revisited [J].ScienceofComputerProgramming.2008, 70(1):1-30.

[8] Satish NarayanaSrirama,PelleJakovits, EeroVainikko. Adapting scientific computing problems to clouds using MapReduce [J]. Future Generations Computer Systems.2012, 28(1):184-192.

[9] Jiang Xiaoping, Li Chenghua, Xiang Wen, et al. Parallel implementing K-means Clustering Algorithm Using MapReduce programming mode [J]. Journal of Huazhong University of Science and Technology.2011, 39(z1):120-124.

[10] Liu Peng. Hadoop in Action - open the shortcut to cloud computing. Beijing: Electronic Industry Press,2011.

[11] Zhou Aiwu, Cui Dandan, PanYong. An Optimization Initial Clustering Center of K-means Clustering Algorithm [J]. Microcomputer& Its Applications. 2011, 30(13):1-3.