International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.21-30
http://dx.doi.org/10.14257/ijdta.2016.9.8.03

# K-means Parallelization Algorithm Based on MapReduce

Shuguang Wang[1] and Chao Jiang [2]

*[1,2]Jilin Communications Polytechnic, Changchun 130012, china*
*[1]Wang_shuguang@sohu.com and 2jch0323@sina.com.cn*

***Abstract***

*Spatial Cluster analysis is another important technique in the field of spatial data mining, especially the K-Means spatial clustering method, which can deal with spatial objects with geographical location and attribute. However, with the development of the information society, the spatial data grows explosively, but the serial algorithm has low computing efficiency and is difficult to process massive spatial data. Aiming at spatial with a double meaning of location and attribute, the paper designed and implemented K-Means spatial clustering parallel algorithm on Hadoop. Using Yahoo Weibo user data is to do clustering analysis. Finally, the visualization of clustering results was implemented by Google Map.*

***Keywords***: *K-Means space clustering algorithm, Hadoop, MapReduce*

## 1. Introduction

Different from common data [1-3], spatial data has non-space features which represent attribute information and spatial features which describe space or location [4-5]. So far, studies on spatial clustering analysis have two types:

(1) From the point of GIS theoretical method and technological tools, most papers made clustering analyses based on the geographical coordinates of spatial objects [6], using object's spatial contiguity as clustering basis instead of the similarity between attributes [7];

(2) Considering from GIS application and geoscientific research, other papers performed clustering analyses of attribute features of spatial objects with the traditional clustering methods [8-9], but ignored the geographical coordinate information of space, not considering objects' spatial proximity [10].

To solve the mining problem of spatial data with geographical location and attribute features, the paper uses K-Means space clustering algorithm to combine geographical location and attributive characters [11], realizing the unification of entity spatial contiguity and attribute similarity. In light of the rapid growth of spatial data scale, K-space clustering parallel algorithm based on MAP is implemented.

## 2. K-Means Spatial Clustering Algorithm

### 2.1. Spatial Clustering

Spatial clustering analysis, as an important branch in the field of clustering analysis study, plays an increasing importance in data processing. The methods used for spatial clustering analysis are mostly traditional clustering analysis technics which are applied for clustering analysis of spatial data. The common clustering algorithms include mainly partitioning method, hierarchy method, density method, grid method and model method. With the exploration and extension of real application and research orientation, a series of new clustering algorithms are consecutively proposed. Those new approaches start from different perspectives and integrate research characteristics in other fields. They each have own unique merits. Of them the representatives include: kernel clustering, spectral clustering, affinity propagation algorithm, quantum clustering algorithm, granularity-based algorithm and intelligent clustering algorithm *etc.* Lots of clustering algorithms are being enriched and improved, which are gradually expanded and extended to the field of spatial clustering analysis field.

Spatial clustering analysis relies on the geometric coordinate or attribute feature information of spatial object entity. Estimate the similarity degree of objects of spatial entity as per certain distance or similarity measuring method. Divide objects into different clusters and make intra-cluster objects possibly similar and inter-cluster objects probably different. It can be solely used as the analysis tool for obtaining the distribution of data objects, analyzing characteristics and rules of cluster conglomeration, and probing into and discover the hidden knowledge therein. On the other hand, it can be regarded as one prerequisite step of pre-processing, such as classification algorithm and feature description method.

### 2.2. K-Means Clustering

K-Means algorithm, and K-means algorithm, called average method, is one of the partition-based classical clustering algorithms which are widely applied. The idea of it is very simple and easily understood. Hence it’s often used in various fields. The basic idea of the algorithm is: initially give randomly or choose one data object as the initial clustering center through one method; traverse in proper order other objects which are waiting for clustering in data collection; calculate the similarity between each object and each determined initial clustering center; follow the principle of the nearest neighbor to add object to the cluster with which it shares the highest similarity; next with the average method re-calculate the clustering center of each cluster as to define a new center; update clustering center and iterate till the clustering center got from previous and posterior iterations won’t change any more or clustering objective function value is smaller than the pre-set threshold value.

### 2.3. K-Means Spatial Clustering

Entity is the basic element of spatial elements. In real world, spatial entity has various features like time features, spatial features and attribute feature. So the formalized model of common spatial entity object is:

$$E = (T, P, A) \tag{1}$$

Where, Time course is the time domain in the time domain $T \in (t_1, t_2, ..., t_n)$, position feature data is $P = (x, y)$, attribute data representation is a tuple $A = (a_1, a_2 ..., a_n)$.The commonest spatial entity model can be hardly used for spatial data mining. Hence it’s necessary to simplify the model for calculation. For the spatial point entity, without regard to time features, the common entity model can be simplified to:

$$E = (x, y : a_1, a_2, ..., a_n) \tag{2}$$

Suppose spatial entity object $P_i$, two-dimension plane coordinate $(x_i, y_i)$ and attribute feature vector $A = (a_{i1,} a_{i2}, ...a_{in})$. The positional distance $P_i$ and $P_j$ ’s attribute distance between any two points in the plane space is calculated as follows:

$$D_p = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$D_p = \sqrt{(a_{il} - x_{jl})^2 + (a_{i2} - a_{j2})^2 + . . . + (a_{in} - a_{jn})^2} \tag{3}$$

Positional distance describes the locational proximity between spatial objects. Attributive distance reflects the degree of similarity between spatial objects. While in spatial clustering, it's required that spatial objects of the same kind approximate spatially and resemble attributively. Traditional K-means clustering algorithms mostly use individually one distance as spatial clustering scale, not able to meet well the above requirements. To make up the shortage, Li Xinyun *et al.* developed three spatial distance calculation equations, which combines geometrical distance and attributive distance:

$$D = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + \sum_{k=1}^{n}(a_{ik} - a_{jk})^2} \tag{4}$$

$$D = \omega_p \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} + \omega_q \sqrt{\sum_{i=1}^{n}(a_{ik} - a_{jk})^2} \tag{5}$$

$$D = \sqrt{\omega_x (x_i - x_j)^2 + \omega_y (y_i - y_j)^2} + \sqrt{\sum_{i=1}^{n}\omega_k (a_{ik} - a_{jk})^2} \tag{6}$$

Considering different effects of spatial domain and attribute domain on the spatial clustering, we think features and discrepancies of spatial objects can only be completely depicted by integrating location distance and attributive distance as for the calculation of the weighted distance.

Hence, K-means spatial clustering algorithm uses spatial entity with the information of spatial position as research object. According to spatial distance measuring criteria of spatial entity objects, it distributes such objects to different clusters. Mean spatial clustering algorithm offers one method which brings together the geographical coordinate and attributive features to spatial distance measure and spatial clustering analysis, which helps to improve the quality of mining information through spatial clustering analyses.

## 3. Design of Spatial Clustering Algorithm Based on K-Mean Value

### 3.1. Parallel Analysis of the Algorithm

Partition-based clustering algorithm is the most widely applied for current scientific research and production practice. K-means algorithm is the most typical clustering algorithm and becomes quickly the breakthrough point of parallel clustering algorithm study owing to its simplicity and practicability. So far, lots of people have proposed MAP-based K-means parallel algorithm and the optimized method. However, the K-means algorithm can't be directly applied for spatial clustering analysis because spatial data themselves have space and non-space characteristics and spatial objects have complicated association. Thus to discuss further, we choose the mean spatial clustering K-means algorithm which realizes the integration of spatial and non-spatial attributes.

It's not hard to notice from the introduction of K-means spatial clustering algorithm that its main calculation is about the spatial distance between each sample and clustering center, as well as that of new clustering center. Calculating spatial distance between them is wholly independent operation, which can be for concurrent calculation; while calculating new clustering center is partially independent. For that we can firstly calculate partial clustering information; next sum all up; and that the first half part can be for parallel calculation. We consider using MAP framework to achieve parallelization of the two parts. In the each iteration, the algorithm performs similar operation. Therefore, MAP-based K-means spatial clustering algorithm can reach the goal by executing the same MAP and RE operations in the each iteration.

In it, to make it easy for MAP parallel calculation model to do data processing, it requires pre-treatment of pending data, making them readable by row as to fetch useful information and use such information as input data of MAP function.

The parallelization process of MAP is consistent with serial thought: firstly randomly choose from pending data K objects as initial clustering center; store them in HDFS as global clustering center to send to sub nodes; then, perform clustering calculation; each calculation includes two processes: MAP and RE.

### 3.1.1. Design of the Function

MAP function's task is to withdraw data object set from pending files; each line in those files represents one object and is expressed in the form of <Key, value>key value pair; then, construct a group of global initial clustering center; calculate the spatial distance between each object in the dataset and each clustering center; allocate data objects to the cluster to which the nearest clustering center belongs, to form a new clustering class and output as <Key, value>key value pair.

MAP process uses all data objects waiting for clustering and the clustering center produced in last iteration as input. In the first iteration, MAP function firstly reads initial clustering center files, which are put like <cluster, feature vector set>. Each line data of sample files waiting for clustering stands for all characteristics of one sample point. MAP function computes the minimal weighted distance between each pending clustering object and all clustering centers, which is assigned to a new clustering. The form of output middle result <Key, value>is (cluster, feature vector set).

### 3.1.2. Design of Reduce Function

The main mission of Reduce function is to realize the process of calculating and updating new clustering center. Reduce process receives intermediate results of MAP function, merges all data objects which have the same clustering K value in each MAP and calculates the average value of each data object feature to have new clustering center and update clustering center file; meanwhile calculates error squared criterion function $E=\sum_{i=1}^{k}\sum_{x\in C_i}|x-x_i|^2$. The form of output key value pair is (mean value of new cluster's intra-cluster objects). Reduce function of the pseudo code is shown below.

Algorithm 1 the pesudop-code of reduce function

```
Input: clustering of the index and the object's feature vector set
Output: the new clustering category, the average value of the cluster object
1.Reduce (Long Writable key, Text value)
{
 2. For (All data objects with the same value)
    {Calculate the average value of each feature;
      Calculate and record the distance from the cluster center;
      According to the above distance, the calculation criterion function
      }
       Output (Clustering category, mean vector)
   }
```

## 4. Experiment Design and Discussion

With the development of technologies in social network, e-commerce, individual recommendation, the number of users grows exponentially. User interaction with the network becomes more and more often. It has been becoming a hot concern to discover user's generic mode from its property, interest and behavior and divide them into different groups based on certain measurement method [12-13]. Yahoo Weibo, as public social platform, has huge user groups which generate enormous user behavior data like user visiting location information, Weibo posting time and contents, user fan's number. How to make clustering analysis through those information to create similar user groups will provide reference basis for community and group analysis, potential geographic friend recommendation, and Weibo advertising and marketing in the social network.

### 4.1. Data Pre-Treatment

Here we use Sina Weibo message data for the experiment, which reaches around 500tens of thousands. Each Weibo includes number, posting time, text, user client, latitude and altitude, author ID *etc.*

In K-means spatial clustering algorithm, object's feature information includes spatial location and attributive information. So it's necessity to filtrate the acquired data. Of them, latitude and altitude reflects user's spatial position information. Fan's number, being focused number, Weibo quantity, and mutual focuses represent user's degree of activity. After deleting some incomplete data, we get 58590 records. It is shown in Table1.

**Table 1. Users Record**

| User id | Latitude | Longitude | Fan number | Being focused number | Weibo quantity | Mutual focuses |
|---|---|---|---|---|---|---|
| 1900456781 | 27.33 | 124.33 | 687 | 896 | 654 | 589 |
| 2374433212 | 27.33 | 124.33 | 256 | 586 | 109 | 186 |
| 2412322214 | 27.34 | 116.45 | 56 | 344 | 46 | 20 |
| 2589066422 | 27.33 | 122.68 | 79 | 80 | 58 | 8 |
| …… | …… | …… | …… | …… | …… | …… |
| 2800020011 | 32.65 | 123.11 | 39 | 209 | 89 | 10 |

### 4.2. Similarity Measurement

Regarding the above user data, user spatial clustering measures the similarity between different users through user visiting location and the degree of activity. Each user information is consisted of three elements, marked User<ID, Position, Activation>. ID is the user identity, Position represents the user access location, Activation refers to user activity.

#### 4.2.1. User Visiting Position

In Yahoo Weibo messages, latitude and altitude information suggests the geographical position of user currently visiting Yahoo Weibo. Users sharing nearer geographic location means they visit much nearer place, which indicates user's similarity in geographic behavior. For instance, two different users may go to the same cinema; then they share common location of watching a movie. With longitudinal and latitudinal data, it's possible to compute the distance between two points on the space [14]. Since two users locate in the region of north latitude and east longitude, the location distance between User(i) and User(j) is:

$$C = \sin(90 - Lat_i) * \sin(90 - Lat_j) * \cos(Lon_i - Lon_j) + \cos(90 - Lat_i) * \cos(90 - Lat_j) \tag{7}$$

$$D_1 = R * Arc\cos(C) * Pi / 180 \tag{8}$$

#### 4.2.2. User Activity Degree

User activity degree [15] refers to how often one user visits one website during some time. The degree implies the rate of user participation in the website, user retention, activity time, days, microblogs, Weibo forwarding and comments are all related with that degree. Those user behaviors are indicators to evaluate user activity degree. Users with similar activity degree have alike behaviors on websites.

Assume that each of the behavioral indicators that reflect user activity is $E_k$ .The weight of each behavior is $W_k$ .Then user activity is expressed as $Activation = (E_1, E_2, ..., E_n)$ . The difference between User(i) and User(j) are:

$$D_p = \sqrt{(E_{i1} - E_{j1})^2 + (E_{i2} - E_{j2})^2 + . . . + (E_{in} - E_{jn})^2} \tag{9}$$

Therefore, the calculation formula of users similarity is as follows:

$$D(i, j) = W_p D_1 + W_k \sqrt{\sum_{k=1}^{n} (E_{ik} - E_{jk})^2} \tag{10}$$

### 4.3. Visualization and Analysis of Clustering Results

Using Google and Android as tools, we annotate and demonstrate clustering results. Figure1 is Google map based on Android.

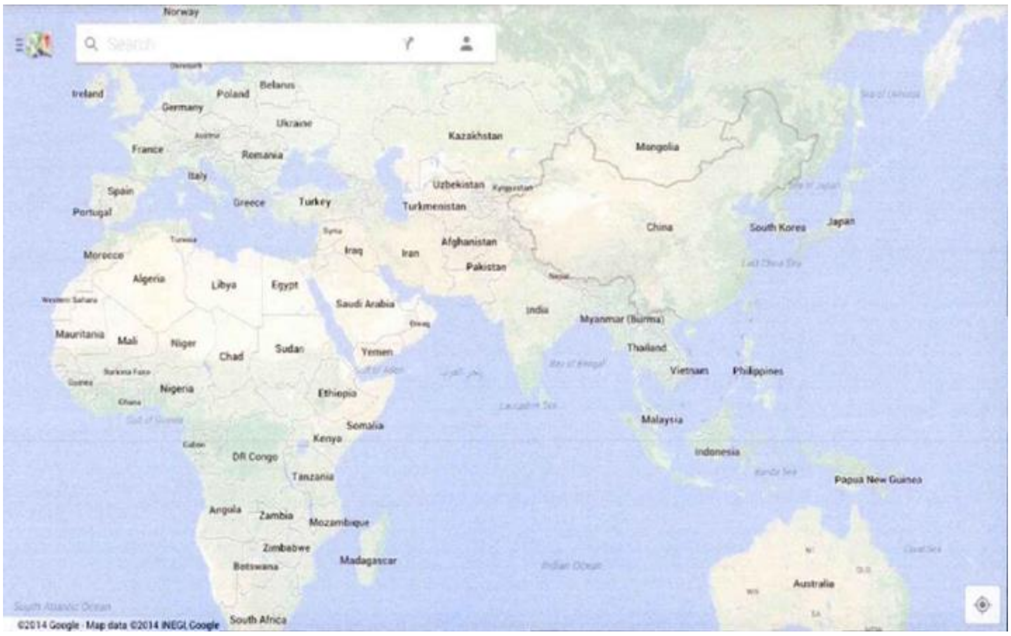


**Figure 1. Google Map Based on Android**

Use user record data as sample data for clustering analysis; each row is a user; totally six dimensions of attributes require calculation. Assume the most iteration times are 500 and the threshold value to terminate the iteration is 0.1. Choose different clustering centers for K spatial clustering calculation; with Google, clustering results can be visualized. Clustering results are shown in Figure 2.

There, different color icons refer to different clustering classes; each icon has user ID. It's not difficult to find that documented Yahoo Weibo users spread mainly in Guangdong, Hainan and surroundings, where the most populated region is Guangzhou and Shenzhen. On the map, what's agglomerated by icons with identical colors are aggregated clusters which are spatially adjacent and attributively similar.

When is $Wp = 1, Wk = 0$ , Clustering according to the position of the space target. Clustering results are shown in Figure 3.

In this case, the attribute information reflecting user activity does not join in calculation; instead, it clusters based on user's latitude and longitude and considers only geographical relationship. Although users in the same class conglomerate into a region, those class clusters do not have practical geographical significances.

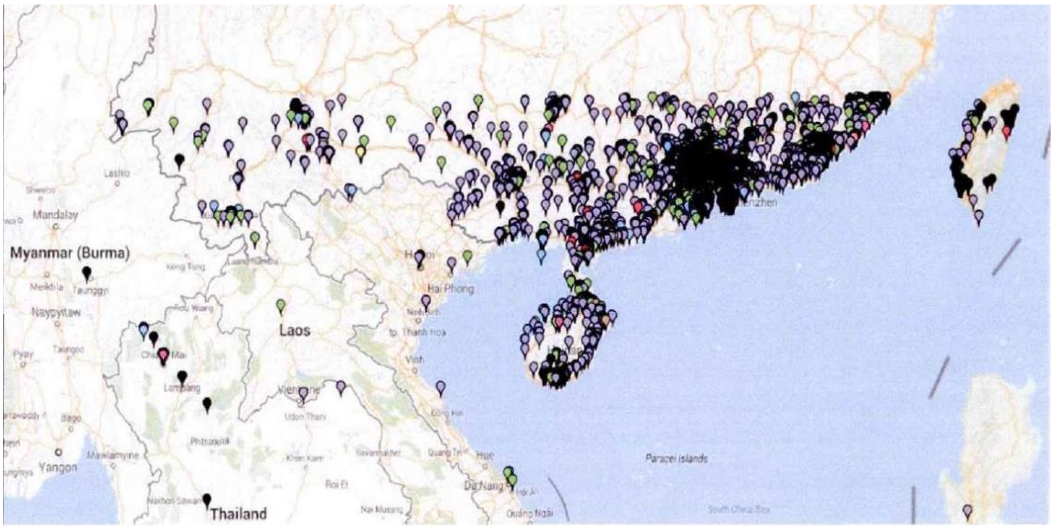


**Figure 2. K-Means Spatial Clustering**

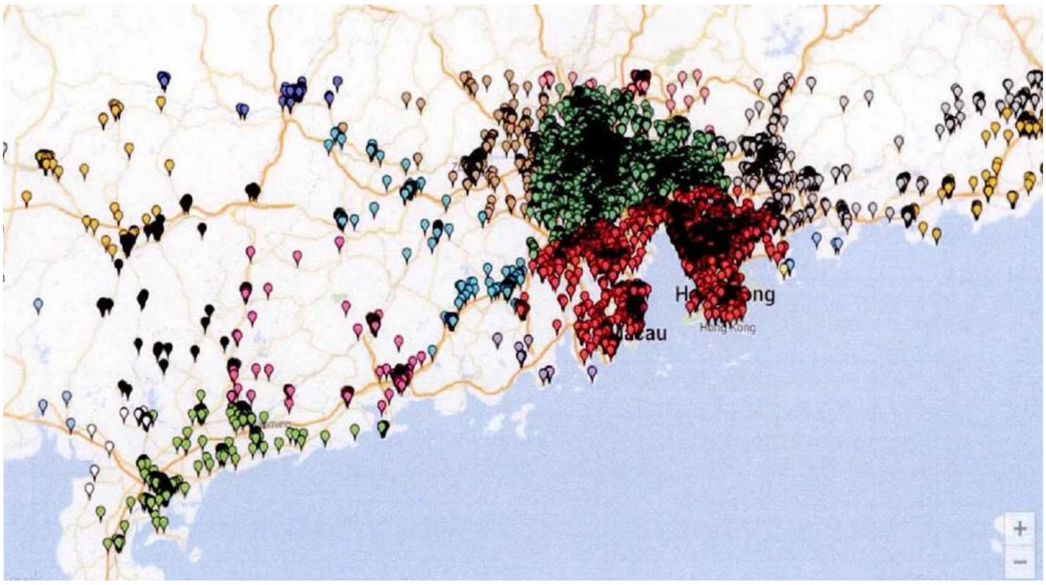
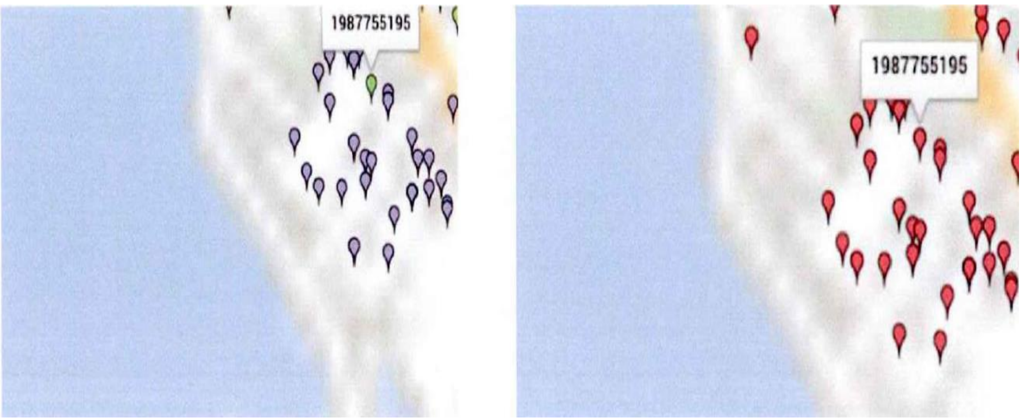


**Figure 3. Position Clustering**

When user activity degree involves in the calculation, in the user group of spatial neighborhood, individuals with greatly different activity degree are excluded of class clusters. Figure 4 is K-means spatial clustering of user at 1987755195, which belongs not to the same class with neighboring users; but in location clustering, they obviously belong to one class.

The user recorded information table is shown in Table2, can be seen as greater difference between the user activity than other users. Considering the degree of Wei and activity of the two factors, the user and the adjacent user are not in the same cluster.



(a) K-Means Spatial Clustering (b) Position Clustering

**Figure 4. The Comparison of Clustering Results**

**Table 2. The List of Users**

| User id | Latitude | Longitude | Fan number | Being focused number | Weibo quantity | Mutual focuses |
|---|---|---|---|---|---|---|
| 1809565655 | 23.33 | 114.33 | 87 | 96 | 567 | 12 |
| 2567800324 | 23.33 | 114.33 | 56 | 860 | 168 | 39 |
| 2789011235 | 23.34 | 114.45 | 6 | 144 | 78 | 178 |
| 3205678995 | 23.33 | 142.68 | 7 | 70 | 189 | 19 |
| 1987755195 | 23.55 | 142.56 | 46 | 12 | 670 | 27 |
| 3367890123 | 25.33 | 143.55 | 789 | 45 | 78 | 0 |
| 2945231131 | 25.333 | 143.55 | 12 | 279 | 450 | 12 |
| 1890212342 | 24.33 | 143.55 | 345 | 112 | 34 | 200 |
| 2765522390 | 24.33 | 143.55 | 39 | 168 | 269 | 0 |

Hence, K-means spatial clustering considers both geographical proximity and activity similarity. From Yahoo Weibo user groups, we find users in adjacent places and with similar activity degree to constitute neighboring similar user groups, offering references to the analyses of societies and communities in social networks. In the era we’re stepping into, users’ interactions with websites become too often and they contain lots of commercial values.

## 5. Conclusion

In the paper, it probed into the problem of multi-layered clustering segmentation of massive image dataset. In normal cases, the quality and speed of image segmentation are contradictory. The algorithm with good segmentation result would work inefficiently; while the efficient algorithm would cause poor precision of segmentation. Sometimes good quality is acquired at the cost of sacrificing speed; and sometimes in turn. The objective of the paper is to consider both quality and speed, for quick and good image segmentation within a certain range.

## Acknowledgement

This work was supported by Research on Education and Teaching Reform of Jilin Vocational and Adult Education.

## References

[1] J. Y. Xie, G. Wenjuan and X. G. Xinbo, “Based on sample space distribution density of initial clustering center of K-means algorithm optimization”, Application Research of computers, vol. 3, **(2012)**, pp. 888-892.

[2] S. Zhongyang, “Remote sensing image classification based on fuzzy C means clustering algorithm of spatial information kernel”, Zhejiang University of Technology, **(2013)**.

[3] L. Wei, “Image segmentation algorithm based on fuzzy C means clustering”, Harbin Engineering University, **(2013)**.

[4] X. Xiaoli, “Research on image segmentation algorithm based on clustering analysis”, Harbin Engineering University, **(2012)**.

[5] Z. Xinye, “Image segmentation method based on clustering analysis”, Dalian Maritime University, **(2012)**.

[6] W. Zonghu, “Research on the key technology of clustering analysis and optimization”, Xi'an Electronic and Science University, **(2012)**.

[7] W. Li, “Knowledge discovery based on rough and fuzzy integration”, Nanjing University, **(2013)**.

[8] X. Yupeng, “Study on spatial clustering analysis”, Harbin University of Science and Technology, **(2015)**.

[9] Z. Jianyu, “Research and application of the mean value of K”, Dalian University of Technology, **(2013)**.

[10] Z. Li, “Global K- means clustering algorithm research and improvement”, Xi'an Electronic and Science University, **(2013)**.

[11] L. Shengxin, “Research and implementation of spatial data clustering analysis algorithm”, Beijing: China University of Geosciences, **(2011)**

[12] C. Kehan, H. long and W. Jian, “Heterogeneous social network based on user clustering recommendation algorithm”, Journal of computer science, vol. 2, **(2013)**, pp. 349-359.

[13] L. Tao and W. Jiandong, “A collaborative filtering recommendation algorithm based on user clustering”, System engineering and electronic technology, vol. 7, **(2007)**, pp. 1178-1182.

[14] S. Yuan, “The similarity of user behavior analysis position of social networking services”, Fujian: Huaqiao University, **(2012)**.

[15] L. Huiji, “Research on user activity in the network community”, Shanghai: Shanghai Normal University, **(2012)**.

## Authors

**Shuguang Wang**, He received his M.S degree from Changchun University of Technology. He is a senior engineer in Jilin Communications Polytechnic. His research interests include data recovery, cloud computing.

**Chao Jiang**, He received his B.S degree from Jilin Animation Institute. He is an engineer in Jilin Communications Polytechnic. His research interests include data recovery, cloud computing.