Group Excavation of Small World Interest Collection Based on LBSN

Junxuan Zhu[1, 2, a], Jianguo Zheng[1]

[1]Glorious Sun School of Business and Administration, Donghua University, Shanghai 200051, China

[2]College of Management Shanghai University of Engineering Science, Shanghai 201620, China

[a]zhujx_2006@ 126.com

Abstract

The continuous development of positioning technology makes the information recommendation based on LBSN become more meaningful. By combining the characteristics of small world network and LBSN, the DBSCAN density clustering algorithm and vector space model are used to calculate the users' similarity based on geographical location. Put forward group mining algorithm by using K-means method and DBSCAN algorithm. Finally, analyzes the real points of interest data obtained from the API of Sina microblogging and gives suggestions on how to achieve accurate information delivery, improve users' experience and create economic value.

Keywords

LBSN,Small world, Interest collection,Data mining.

1. Introduction

The emergence of large data age has promoted the development of location social network (LBSN), which not only has the characteristics of social networks in the past, but also contains a lot of location information and interest information [1].

In the social network platform, users can share the experience through the comments and can also add friends or become a fan of others to establish contact, these two characteristics are the basis of LBSN [2].The information of users, in the LBSN, check or comment not only recorded the user's behavior and hobbies, but also to provide the researchers an opportunity to study their behavior [3]. Foreign researchers have integrated large LBSN data sets into different social networks (using snowball sampling) time or spatial granularity to test the statistical model, noting that LBSN data is well suited to research issues and perspectives on social or spatial phenomena [4]. With the rapid development of LBSN, POI recommendation also provides an excellent opportunity for LBS [5]. A point of interest recommendation is a personalized recommendation based on contextual information which is an important way to help people find interesting places [6]. However, in real life, the user in the LBSN is often sign in a small part of the interest points, which led to the user sign the scene information and historical data is extremely sparse [7], and greatly increased the difficulty of the recommended.

Make a data mining of the user community on the small and medium-sized network of LBSN on the basis of the similarity research between the user groups in the geographical position and the characteristics of the small world network and LBSN. Then use the small-world network model to classify the users. Finally analyzing the real point of interest data from the Sina micro blogging mobile phone, founding the characteristics of micro blogging community based on LBSN and giving advises in several ways, such as how to achieve accurate information delivery, improve the user experience and create economic value.

2. User population similarity calculation based on geographic location feature in LBSN

2.1 Geographical location analysis

Geographical location characteristic is the characteristic of the geographical location, including the specific latitude and longitude information of the location which mainly refers to the relevant information of its land and sea position, hemisphere position, latitude and longitude range in geography [8]. Here, the geographic location feature mainly refers to the latitude and longitude information of the location, the type of the point of interest, the city name, the address, the name of the place, the number of the user signs, the number of uploaded photos and so on. Compared with the specific features of the landform, cultural information is more valuable in the data mining and can reflect the user's preferences and interests better. Therefore, the location and geographical features are defined as the latitude and longitude information, POI, address, the number of photos uploaded and other relevant cultural information of the specific location.

2.2 User population similarity calculation model based on geographic location feature

2.2.1. Hierarchical clustering based on POI

In LBSN, geo-location features include POIs. Generally speaking, the attributes of POI can be expressed as POI = {ID, Name, longitude and latitude} [9]. This access record can be represented by a quaternion Register = {User ID, POI, Time, Rate} when the user arrives at a location, signs in the social network and records the access record [10].

Because there are much different POIs around a certain geographical location. Therefore, DBSCAN method was chose to do density-based clustering of adjacent POI. DBSCAN is a classic density-based clustering algorithm that can effectively identify noise points. And it can also identify and filter POI that are accessed only by very few users or those places with very few user access frequencies [11].

The specific flow of the DBSCAN algorithm is as follows:

Input: POI set $P = \{p_1, p_2, ..., p_n\}$, neighborhood radius E, density threshold MinPts;

Output: All SOI sets $S = \{S_1, S_2, ..., S_n\}$ of density reachable conditions after clustering

(1)To detect the object p in the test set, if it is in the state of $unvisited$ (not classified as a cluster or marked as a noise point), checking whether its neighborhood is greater than $MinPts$. If not satisfied, to create a new cluster S and all of them add to the candidate collection T.

(2) All $p \in T$ in the state of $unvisited$, checking its neighborhood, if the number of objects is greater than or equal to $MinPts$, then these objects will be added to T, if p didn’t classified into any cluster (SOI), it will be brought into T.

(3) Repeat step (2) to continue checking the unprocessed object until the set T is empty.

(4) Repeat steps (1) - (3) until all points are listed in the cluster (SOI) or are marked as noise.

So, define SOI (Set of Interest) as a set of POIs clustered by DBSCAN algorithm. It should be noted that the DBSCAN algorithm needs to preset two parameters: the neighborhood E and the number of points that can be included in neighborhood $MinPts$. In general, the closer the location where the users arrive is, the more similar the mode of action is.

2.2.2. Similarity calculation method and model construction based on geographic location feature in LBSN

Combining the content mentioned above, the user's trajectory shows the user's moving habit of geographical location in the real world. So we stipulate that: (1) the closer the location of the user's visit is, the more similar the action trail is. (2) The more users access the similar geographical location, the higher the user's similarity is.

On the basis of above rules, calculating the user similarity by combined with the vector space model. By cluster analysis, the SOI of each user's check-in position is classified as a vector $\vec{A}=[a_1,a_2,...,a_n]$ and the number of times the user visits the same SOI is a_i . The more the user visits the same SOI, the higher similarity the users' action is. And the SOI of the user visited constitute a user access location matrix $V_{m\times n}$.

$$V_{l(m\times n)}=\begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n-1} & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n-1} & v_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn-1} & v_{mn} \end{pmatrix}$$

The number of users is m, the number of SOI in neighborhood E is n, the number of the ith user visits the jth SOI is v_{ij} and the number of layers by dividing clustering hierarchies is l.

The vector at the n-dimensional space position is used to represent the user's position and the similarity between the users is represented by the value of $\cos$ between the vectors. Assuming that the user A and user B are represented by vector $\vec{A}$ and $\vec{B}$ in n-dimensional space, the similarity calculation method between user A and user B is:

$$sim(A,B)=\cos(\vec{A},\vec{B})=\frac{\vec{A}\vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

The POI is clustered in each neighborhood E, the similarity in different neighborhoods and the similarity of the user population are calculated. Among them, the user similarity of the cross-level clustering can be expressed as: $Sim_{overall}=\sum_{i=1}^{F}\mu Sim_i$, $\mu=\frac{\beta_i}{\sum_{i=1}^{H}\beta_i}$, F is the total number of levels, Sim_i is the user similarity on the ith layer, β_i is similarity weight of the ith layer, the closer the SOI of users in the higher level is, the greater the weight will be.

2.2.2. User similarity calculation step

Calculating the similarity of user's geographical location mainly consists of the following steps:

Step 1: determining the two parameters of the $DBSCAN$ algorithm: the neighborhood E and the density threshold $Minpts$, according to the geographical characteristics of the POI.

Step 2: cluster the POI multiple many times with different neighborhood parameters E (scanning radius) and $Minpts$ (density threshold);

Step 3: build the user's position matrix by combining the status of SOI that users accessed on each layer.

Step 4: Calculate the similarity of the user on each layer;

Step 5: According to the different weights of each layer, the overall similarity between users is calculated by the similarity of the users at each level.

3. Community groups mining of users' interested collection in small world network in LBSN

3.1 Definition of user community in small community network

Small world network model as a the promotion of the six-degree separation theory, its effect can be summarized by the theory of Six degree segmentation, refers to an arbitrary information can be transmitted to another stranger through information transfer between 6 acquaintances [12]. The characteristics of the small world network can be summarized as the length of the average path in the network is very short and the network clustering factor is quite high.

Based on the characteristics of the small-world network and the community structure, we define the user community in small-world network as users who are in different networks with the very small average network path and high clustering factor. Such user community has the characteristics of the community structure and the small world network.

3.2 Introduction of the method to discover user community group in small world network

What called founding community in small world network is the process of dividing the small world network into usable and meaningful societies. It's essentially a clustering problem, a typical and common unsupervised learning method in machine learning. Therefore, choose the most classical clustering algorithm- K-means which based on partitioning method.

The main content of the K-means method is to select K points as the initial barycenters and assign each point to the barycenter closest to them, then form K clusters and recalculate the barycenter of each cluster. The points other than the initial barycenters are assigned to these new centroidings and continuous iteration update until the cluster does not change or the algorithm reaches the maximum number of iterations, and then stop the algorithm.

The spatial complexity of the algorithm is $O\left(\left(m+k\right)n\right)$; the time complexity is $O\left(tKmn\right)$, where K is the number of initial centroids, m is the number of records, n is the dimension, t is the number of iterations [14].

3.3 Community mining algorithm based on LBSN small and medium-sizednetwork users' interest set

In LBSN, the user's check-in information, which is a very important part of the geographical location, can be reflected by the user's latitude and longitude coordinates. According to the different data types, choosing the appropriate clustering algorithm is the basis of the reliable results. In the construction of the community mining model, the K-means method and the DBSCAN algorithm are feasible and each has its advantage. Therefore, we choose the K-means method and refers to some ideas in the DBSCAN algorithm.

Specific flow diagram is shown in Figure 1.

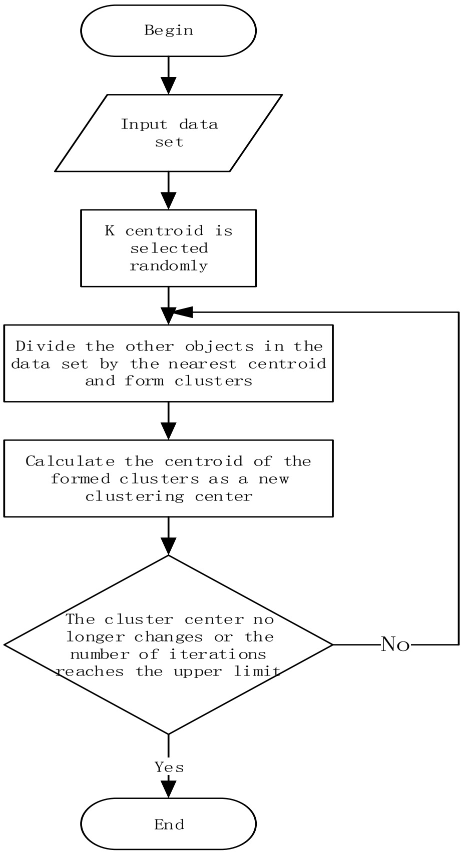

Begin
Input data set
K centroid is selected randomly
Divide the other objects in the data set by the nearest centroid and form clusters
Calculate the centroid of the formed clusters as a new clustering center
The cluster center no longer changes or the number of iterations reaches the upper limit
No
Yes
End

Fig. 1 Algorithm flow diagram

4. An Empirical Analysis of Community Group Mining Based on LBSN

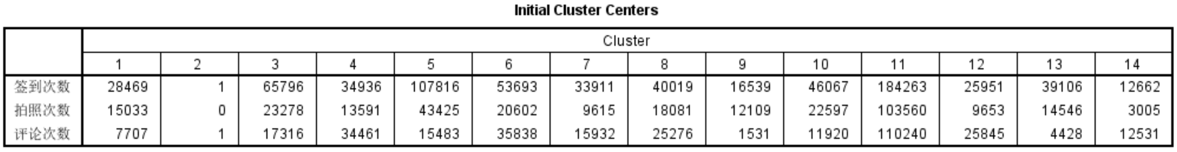
4.1 Data sources

The data set is the database of the Shanghai Sina micro blogging POI data that downloaded from the database of CSDN Forum. The members of the Forum crawled data from the public API of Sina micro blogging, the total number of data entries is 284118. The data table includes POI name, POI address, POI type, POI longitude and latitude, check number, number of taking photoes, the number of comments and other major information of interest point. There is difference between the datas. Because the number of data and the type of POI are large, it’s more difficult to see the potential information from the data, in line with the conditions of data mining and the subject of research. So it can be used as experimental data.

In the data cleaning, firstly screen out a large number of interest points whose check-in value is 0. Secondly, because the classification of POI in the original data is more close (more than 200 categories). So do classification once again, then these subcategories will be classified into 14 categories which can improve the reliability of the data.

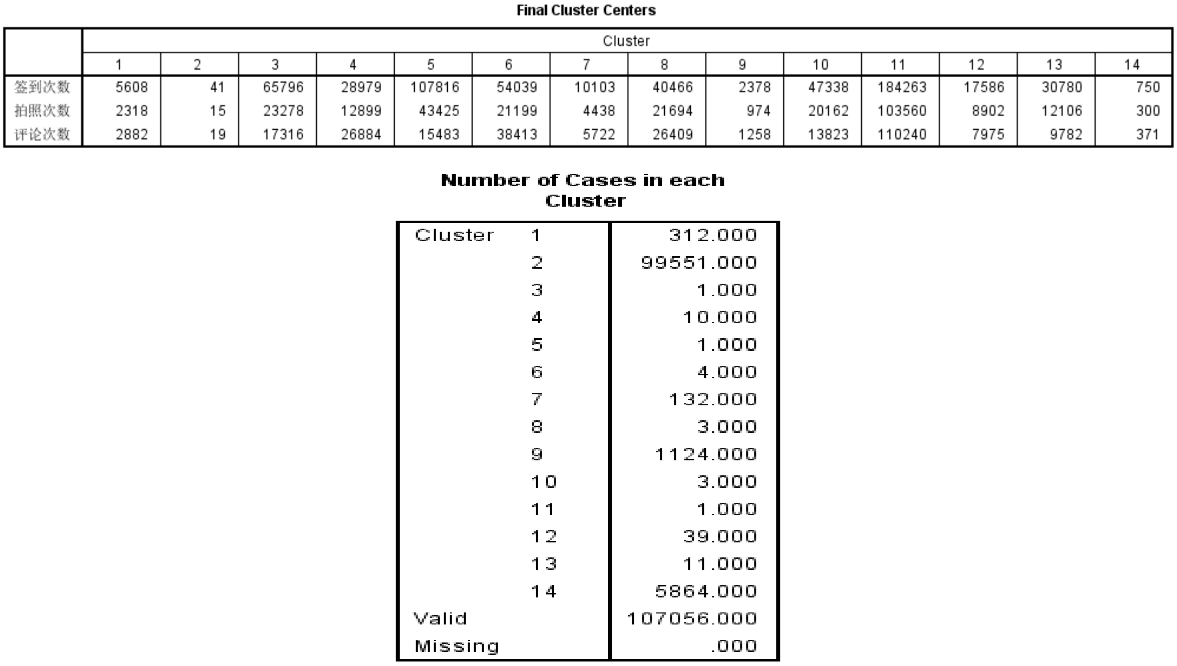
4.2 Experimental results

The processed data set is imported into the SPSS program for K-means analysis. Since the POI type is divided into 14 classes during the initial processing of the data, the K value is set to 14 here and the number of iterations is set to the maximum value of 999. The results are shown in Figure 2 below:



Initial Cluster Centers

	Cluster													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
签到次数	28469	1	65796	34936	107816	53693	33911	40019	16539	46067	184263	25951	39106	12662
拍照次数	15033	0	23278	13591	43425	20602	9615	18081	12109	22597	103560	9653	14546	3005
评论次数	7707	1	17316	34461	15483	35838	15932	25276	1531	11920	110240	25845	4428	12531



Final Cluster Centers

	Cluster													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
签到次数	5608	41	65796	28979	107816	54039	10103	40466	2378	47338	184263	17586	30780	750
拍照次数	2318	15	23278	12899	43425	21199	4438	21694	974	20162	103560	8902	12106	300
评论次数	2882	19	17316	26884	15483	38413	5722	26409	1258	13823	110240	7975	9782	371

Number of Cases in each Cluster

Cluster	1	312.000
	2	99551.000
	3	1.000
	4	10.000
	5	1.000
	6	4.000
	7	132.000
	8	3.000
	9	1124.000
	10	3.000
	11	1.000
	12	39.000
	13	11.000
	14	5864.000
Valid		107056.000
Missing		.000

Fig. 2. The first experimental results

In the first experiment, the number of iterations was 110. Observing the final cluster center and the number of POI included after clustering, it can be found that POIs are clearly clustered in the second cluster. In addition, the four clusters of 1, 7, 9 and 14 has more POI aggregation than the other 10 clusters which included small number of POIs. So we turn the value of K from 14 to 5 for the second control experiment.

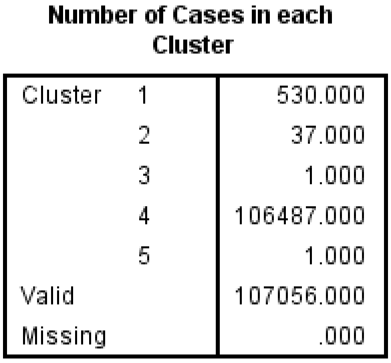
In the second experiment, the K-means algorithm has 19 iterations, and the initial clustering center, the final clustering center and the number of POIs included in each class is shown in Fig.3.

Initial Cluster centers

	cluster				
	1	2	3	4	5
sign-in	39106	34936	107816	1	184263
take photos	14546	13591	43412	4	103560
comment	4428	34461	15486	1	110240

Final Cluster centers

	Cluster				
	1	2	3	4	5
sign-in	7286	34540	107816	102	184263
take photos	3174	14952	43412	40	103560
comment	3792	19835	15483	51	110240



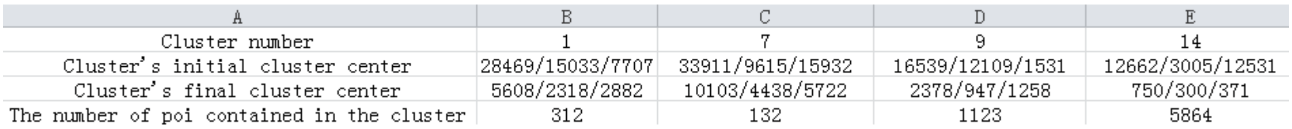
Number of Cases in each Cluster

Cluster	1	530.000
	2	37.000
	3	1.000
	4	106487.000
	5	1.000
Valid		107056.000
Missing		.000

Fig. 3. The second experiment results

According to the results of the two experiments, it can be founded that some of the fewer POIs with fewer check-in occupy the majority of signings, such as the number of POIs in the second cluster is 99551, the average number of check-in, the number of taking photographs, the number of comments respectively were 41, 15, 19. The characteristics of these POIs are visited by a large number of visitors and the number of checkpoints and the number of evaluations was larger than others, but the users will not to visit the same place many times.

Because the K-means algorithm is very sensitive to the outliers, these maxima have some effect on the algorithm after the initial data processing. But because these maxima are very few, these POIs can be classified by simple filtering. It is worthy noting that after removing the second cluster and the maximum value, there are still four types of clusters containing more POIs. The number of POIs in these clusters is taken as follows:



A	B	C	D	E
Cluster number	1	7	9	14
Cluster’s initial cluster center	28469/15033/7707	33911/9615/15932	16539/12109/1531	12662/3005/12531
Cluster’s final cluster center	5608/2318/2882	10103/4438/5722	2378/947/1258	750/300/371
The number of poi contained in the cluster	312	132	1123	5864

Fig. 4 Class 4 contains more clusters of elements

Then, a brief analysis of the four clusters of 1,7, 9,14 is made in order of the number of POI.

The cluster with a large number of POIs is 14, which contains 5,864 POIs and the number of check-in, take photos and comment are 750/300/371. By observing the number of check data in dataset were 247 between 730-770 (± 4%) and can found that these types of POI does not include tourist attractions and government agencies, which include the main types of POI are residential buildings (29.96%), catering (20.65%), others (19.84%) and transport services (12.14%). Combined with clustering results and data analysis can get the community's general characteristics. The community not only innclude some crowded community and dining venues, but also contains some personnel-intensive transport hub and other miscellaneous. This community is characterized by more intensive staff. Because the data is obtained from the public API of Sina Weibo, so we can also infer that there are many microblogging active users in this community, that is the Post-80s and the Post-90s.

The second largest POI cluster is No. 9, which contains a total of 1123 POIs and the sample range of checkin data is [2259, 2497], 127 in total. It does not contain the type of government agency POI, which includes the main POI type: food and beverage (16.54%), others (14.96%), transportation services (11.81%), residential buildings (11.02%), life entertainment (10.24%), convenience facilities (10.24%), culture and education (9.45%), accounting for 84.26% of the total. If consider the main types of shopping services (6.30%) that are not included, these seven POIs account for 90% of the total. Compared with No. 14, there are fewer residential areas in No. 9, the number of sign, upload photos and reviews are larger and the unit residential area can get more services. The higher number of signings indicates that the area has a large flow of people and the lower proportion of the occupations indicates that the population in the region is large but the number of households is not so much. It can confirm this speculation to some extent.

The third group of POI cluster is No. 1, which does not include public facilities, government agencies and fitness sites. The number of check-in times range is [5385, 5832], a total of 49 datas. The main POI types are: entertainment (30.61%), culture and education (16.33%), shopping service (16.33%), food (18.37%), four total 81.64%, of which convenience facilities, residential buildings are only 2 (4.08%). It can be seen that the characteristics of this community are very different with No. 14. The main features of this community are that the residential buildings and the corresponding supporting convenience facilities are few, but food and beverage, shopping services, life and entertainment, culture and education are intensive. So it can be speculated that the community is mainly located in the main shopping and catering (similar to the Nanjing Road Pedestrian Street) or student-intensive campus street along the street shops (similar to Song Jiang University City). At the same time, there are many active microblogging users in this region who are willing to sign in, so it can be inferred that this area is famous shopping / dining / entertainment / education block in Shanghai.

The last number is NO.7. Because the included POI is less than the other three clusters, both the ratio of the initial cluster center and the ratio of the final cluster center is much larger than the other three clusters. So the range of the data is adjusted to ensure that sample data extracted to meet the demand. The numerical range is (9104, 11122), a total of 46 which not include two types POI of commercial land and government agencies, the main POI types were: shopping service (23.91%), food and beverage (14.29%). The main types of POI are medical related (10.87%), life entertainment (8.70%), culture and education (8.70%), public facilities (8.70%) and the remaining POI types (6.52%). It can be seen that the NO.7 is similar NO.1, but NO.7 consists of shopping service primarily and catering secondly. There are more POIs that is related to medical. Compared with the previous analysis of the NO.1 and NO.14, although the NO.7 listed the six main types, but the total percentage is only 75.11%. The percentage of shopping service is more than others and the distribution of other types are closer. And combined with the information that residential buildings are few, you can see the elements of the community is very rich.

5. Conclusion

By combining the characteristics of small world network and LBSN, put forward the group mining algorithm to understand the user's demand while calculating the similarity degree of users based on the geographic location. To achieve the goals of interested information putted, user experience improved and economic values increased. In the accurate information delivery, through the data mining process to understand the daily habits of users and know the frequency of user activities in different places. The community related information can be putted through the relationship between community and geographic position and can also be putted through intensive degree of combined with the intensive drill of the user, interested points in the geographical location which can effectively achieve the goal. As for improving user experience, link the check in behavior with other behavior to guide the user to sign in which will make the data be more fully and improved the user experience. In the creation of economic value, through the analysis of information in the LBSN to understand the user's interest in different POI and to make reasonable information recommended to obtain economic benefits.

References

[1] Li Xin, Liu Guiquan, Li Lin, Wu Zongda, Ding Junmei.Application Algorithm of Social Relations Mining Based on Interest Circle in LBSN [J] .Journal of Computer Research and Development, 2017, (02): 394-404.

[2] Qu Hongyang, Yu Zhiwen, TIAN Miao, GUO Bin.Study and Implementation of Commercial Site Selection Recommendation System Based on LBSN [J] .Computer Science, 2015, (09): 33-36 + 44.

[3] Clio Andris. LBSN Data and the Social Butterfly Effect[J]. Acm Sigspatial International Workshop on Location-based Social Networks,2015:3

[4] Xu Zening, Gao Xiaolu. Construction method of urban built-up area boundary based on electronic map points of interest [J]. Acta Geographica Sinica, 2016, (06): 928-939.

[5] Zhang Tieying, Li Hongwei, Xu Chao, Meng Chao, Zhu Yan. Method of interest point data visualization using density clustering algorithm [J]. Journal of Surveying and Mapping Science, 2016, (05): 157-162.

[6] Gao Rong, Li Jing, Du Bo, Yu Yonghong, Song Chengfang, Ding Yonggang. A position that integrates situational and critical information. Social network interest point recommendation model [J]. computer research and development, 2016, (04): 752-763.

[7] Zhu Lichao, Li Zhijun, Jiang Shouxu. Survey of location-based social networks [J]. intelligent computers and applications,.2014, 4 (4): 60-67).

[8] Qu Li, Yu Zhiwen, Tian Miao, Guo Bin. Research and implementation of recommender system for commercial location based on [J].. LBSN computer science, 2015, (09): 33-36+44.

[9] Qu Shaoling, Hu Dehua. Visualization Analysis of Small World Theory [J]. Library Journal, 2016, (06): 57-65.

[10] Song Jianlin. K-means clustering algorithm to improve [D]. Anhui University, 2016.

[11] Wu Qingxia, Zhou ya, Wen Di Yao, he is red. The user interest and interest of popular tourist routes of personalized recommendation [J]. based on computer applications, 2016, (06): 1762-1766.

[12] Xiong Shuming, Hu Yongdi. Topological control of heterogeneous sensor networks based on small world concept [J]. Computer Engineering and Design, 2016, (11): 2869-2875.

[13] Liu Na new, Wu Zhongxin, Lin Jianfeng.Study on the risk transmission process of small and medium-sized enterprises in science and technology - based on the perspective of small world network [J]. Accounting Research, 2015, (01): 56-60 + 97.

[14] Zhang Feng, Xie Zhenhua, Jiang Tao, Cui Xu Hengbo. Pearson Kao Lun, a combination of vector correlation coefficient weighting method based on [J]. command and fire control, 2015, (05): 83-86.