Communication Base Station Log Analysis Based on Hierarchical Clustering

Shao-Hua ZHANG and Chang-Hua LIU[a]

School of Mathematics & Computer Science, Wuhan Polytechnic University, Wuhan, Hubei, China

Abstract. Communication base stations generate massive data every day, these base station logs play an important value in mining of the business circles. This paper use data mining technology and hierarchical clustering algorithm to group the scope of business circle for the base station by recording the data of these base stations. Through analyzing the data of different business circle based on feature extraction and comparing different business circle category characteristics, which can choose a suitable area for operators of commercial marketing.

1 Introduction

Today mobile terminal penetration and usage occupy a high proportion in People's Daily life. With the development of the Internet, the mobile network in urban and rural areas in our country basically achieves the all-weather and full coverage[1]. The study found that although mobile phone positioning data is discrete and sparse, but the mobile phone data can still be people's activities for high-precision prediction[2]. This conclusion provides a theoretical basis for the study of people's activities. The user's mobile phone positioning data to some extent can also reflect the user's habits. Real-time communication between the mobile terminal and the base station will generate the base station cell number information of the mass-tagged mobile terminal time series, which will be uploaded to the server terminal for storage[3]. Business circle represents the gathering area of the people's production and consumption. According to the coverage of mobile phone signal in real geographical space, the real-time location data of mobile phone users can be mapped to the corresponding real geospatial location, which can restore the realistic trajectory of users realistically and objectively. Differentiating the business scope of the current area and pushing for accurate business activities by excavated the link between regional population distribution and activity characteristics. The analysis of mining high-value information log has an important practical value[4]. Through precision marketing, businesses can get huge profits. So more and more researchers have devoted to the studying of the base station log.

Data mining is to extract the hidden information from the massive existing data by the algorithms[5]. Nowadays, data mining is applied in many fields. Data mining is one of the most important applications in the field of data mining. Domestic and foreign scholars have done a lot of research in this field. The log information in the communication base station covers a wide range of people, high real-time data and long acquisition period. It can record the spatial and temporal

[a] Corresponding author: liuch@whpu.edu.cn

characteristics of urban residents' activities for a long time, and there is no pressure on the respondents, so the academics think that it can dig out the characteristics of resident activities on the current area efficiently[6].

From the mass mobile phone positioning data of base station to obtain the characteristics of the stream of people, which can lock the current area of population activities effectively and found a high value area quickly to target business promotion. In the existing literatures, they analyze the population flow types according to the log of base station to find out the region type. For example, Zhigang Zhang and others through the design of two kinds of distributed mining algorithms(GPMA and SPMA) to dig out the important position[7]. A general framework is proposed to improve the usability of trajectory data in his paper, including a filter to improve data usability and a model to get the mining results. Xin yang Kong, etc. analysis the population movements through the using of SD-JUPF decision algorithm for the mobile phone log[8]. In their paper, a Movement Features-based Judging Urban population Flow(MF-JUPF) algorithm utiliziing cellphone trajectory data was proposed to deal with the is sure about the population flow. Jia Chen, etc. using the DBSCAN method to carry on the extraction to the mobile phone location data and identify the characteristics of the users[9]. In order to acquire personal profiles they propose a resonable technical route that first extracts the geographic regions from personal mobile phone location data based on a density-based clustering algorithm.

In this paper, the calculation model of the human flow characteristic index is established, and the calculated index value is clustered by the hierarchical clustering algorithm. To analyze the experiment data through part of the base station provided by the local mobile operators. The results show that the algorithm can achieve good clustering effect, and can be divided into different types of business district, which shows the feasibility and effectiveness of the algorithm in data mining. This provides a reference for accurate business marketing. Therefore, this paper studies the base station log, proposes the use of hierarchical clustering algorithm to find out the important information, which can provide a new idea for future research work.

2 Business circle division method

The use of mobile phones' related services such as text messages, browsing the Web, switching machines, etc. will generate log records. The current mobile phone users connected by the base station number, connection time, and the information such as user EMASI number[10] can identified by these records. The activities of the corresponding mobile phone users are reflected in log records. Each of base stations' service area on behalf of the different business circle type. Clustering analysis of mobile phone users within the coverage area of the base station, summarizing their demo graphic characteristics, and identifying different types of base station coverage areas, is equivalent to identifying the different types of business circle.

2.1 Analysis of activity area category

This paper extracts the characteristics of people flow in the coverage area of the base station and then finds out the high-value business circle by the base station. Due to the coarse-time and coarse-grained limits of mobile phone base station data, The active areas are identified by the users' activity time[11]. High-value business circles has the characteristics of large flow and long percapita residence time. However, due to office workers working in the base station during the day in which the location is fixed and the residence time is relatively long. At the same time, residential area residents in the base station range is basically fixed, residence time is relatively long. It is difficult to distinguish the high-value shopping district, residence and work area only through the retention time as a flow characteristics. Therefore, the extracted features of the flow must be able to more clearly distinguish between these base stations coverage. The following assumptions also needed to divided the flow characteristics in this paper:

1)Only one active area is the users' residence

2)The user has at least one workspace

3)The user can only work in the region and can not appear in other regions in working hours.

According to the above suppose, we can use the residence time per person in working hours, percapita residence time in the early morning, percapita residence time at the weekend and daily traffic as the base station coverage area of the flow characteristics. The users' activities are divided into the following: early morning hours (00:00~07:00), work hours (09:00~18:00).On the other hand, due to the fact that the working time and the daily life of some people are different, the premise of this paper may not be applicabled to all users in real life, and the corresponding rules need to be reformulated for special activities. After analysis and discussion, we divided four types of active regions. There are four flow characteristics as follows:

1)residence time per person in working hours: All users during working time (09:00 ~ 18:00) in average time within the scope of the base station. This indicator can be used to characterize the flow characteristics near the base station during the working period.

2)percapita residence time in the early morning: It represents the average residence time of people in the base station service area in the morning(00:00 to 07:00).Generally in this time period, the users are in the rest, through this indicator can be excavated pedestrian flow characteristics of residential base stations.

3)percapita residence time at the weekend: It stands for people around the base station per capita residence time at the weekend. High value business circle the number of shopping at the weekend and time will be greatly increased. With this indicator can indicate the flow characteristics of high-value shopping district.

4)daily traffic: It represents the number of people in the service area of the base station everyday. The more people, the greater the possibility of consumption. So it can reflect the characteristics of the high value business circle of people.

2.2 Construction of the flow characteristics of the calculation model

Supposed the total length of the user's date in the raw data is S days. The total number of base stations that all users pass through are X, the number of users are Y. The m on behalf of a day and n represents a user. Determine the current user's residence time by calculating the time difference between the two records. Then for a base station x, the time that the user y stays on the day of the week is w_t_{xy} . A user's retention time is n_t_{xy} in the morning and his(her) retention time is z_t_{xy} over the weekend. A user retention time is d_t_{xy} a day. If a user is not in service area for the base station, the four indicators of 0. For a base station, the base station coverage area of the flow characteristics of the formulas are as follows:

1)residence time per person in working hours

$$w = \frac{1}{SY}\sum_{m=1}^{S}\sum_{y=1}^{Y} w\_t_{xy} \tag{1}$$

2)percapita residence time in the early morning

$$n = \frac{1}{SY}\sum_{m=1}^{S}\sum_{y=1}^{Y} n\_t_{xy} \tag{2}$$

3)percapita residence time at the weekend

$$z = \frac{1}{SY}\sum_{m=1}^{S}\sum_{y=1}^{Y} z\_t_{xy} \tag{3}$$

4)daily traffic

$$d = \frac{1}{S}\sum_{m=1}^{S}\sum_{y=1}^{Y} d\_t_{xy} \tag{4}$$

2.3 Hierarchical clustering algorithm

Clustering is the process of grouping physical or abstract collections into multiple classes of similar objects. Clustering analysis is a group of data according to the similarity and differences into several categories. The purpose is to make the similarity between the data of the same category as large as possible, and the similarity between the data in different categories is as small as possible[12, 13]. Hierarchical clustering algorithm for a given level of decomposition of the data until a certain condition is met, can be divided into specific types of cohesion and split. Cohesive hierarchical clustering is a bottom-up strategy. At first, each object is used as a cluster and then the cluster is combined into more and more large clusters, until all the objects in a cluster or a condition is terminated. The split level clustering is just the opposite of the hierarchical clustering. Split hierarchical clustering uses a top-down strategy. First, it puts all objects in a cluster and then subdivides them into smaller and smaller clusters until each object becomes a cluster or to a termination condition[14]. The basic idea of hierarchical clustering algorithm is to calculate the distance between each object, and the closer object merged into a class. Loop execution until merged into one large class[15]. In this paper, a hierarchical clustering algorithm is used. Hierarchical clustering algorithm includes the following four steps:

Step1: Each object is classified as a class, so the total number of categories is N. Each category contains only one object. We should calculate all the categories of the distance between each other.

Step2: Find the two categories with the smallest distance and merge it into one class.

Step3: Recompute the distance between the new category and all the old categories.

Step4: Repeat Step2 and Step3 until the last merge into one class.

3 Experiment and analysis

3.1 Experiment environment

In this paper, the Matlab R2014b is useded to realize programming. Experiment environment as follows:

Processor: Inter (R) Core(TM)
CPU: i3-2330M
Frequency: 2.20GHz
Memory: 4G
Operating System: Windows 7-64 bit

3.2 Experimental data preprocessing

In this paper, the positioning data obtained from the analysis of the specific interface provided by the local mobile communication operators. Select the base station data from 2014-1-1 to 2014-6-30 as the observation window for analysis. Selecting the base station data from 2014-1-1 to 2014-6-30 as the observation window of the analysis and extracting the location data of the city to form the modeling data. The partial position data of the base station of the period is shown in Table 3-1. There are two types of network: 2G and 3G in the table. The LOC number represents the base station location number. The EMASI represents the base station location number.

Table 3-1 Part of the base station location data

Date	Time	Base Station Number	EMASI Number
2014-1-1	02:21:32	36983	55555
2014-1-1	02:25:43	36981	55555
2014-1-1	02:35:13	36981	55555
2014-1-1	02:55:23	36981	55555

By analyzing the base station mobile log found that part of the log records incomplete, we call these data "noise"[16]. On the one hand, due to the huge amount of data, and the log contains a lot of "noise" data, which is needed to denoising. On the other hand, there are many attributes of the positioning data, which need to filter out the useless attributes. So we need to clean the data in the first step. There are three types of network type. The LOC number and signaling type are useless for the purpose of this paper, and the three kinds of redundant attributes are eliminated by word segmentation tools. Measuring the user's retention time does not need to be accurate to the millisecond, so the time will be removed in the same way. After the above steps on the collation of the log, we get the log after the pretreatment. As we can see, the Table 3-2 shows the format of data pre-processing.

Table 3-2 Preprocessing data

Date	Time	Network Type	LOC Number	Station Number	EMASI Number	Signaling Ty pe
2014-1-1	02:21:32:45	2	99598990383092	36983	55555	334109CA
2014-1-1	02:25:43:23	3	93592190333561	36981	55555	334204CA
2014-1-1	02:35:13:03	2	93592190333561	36981	55555	334204CA
2014-1-1	02:55:23:00	3	93592190333561	36981	55555	334204CA

3.3 Experimental results

By calculating the switching time between successive two base stations (by subtracting the last release time of the last base station from the initial connection time of the next base station) to represent the dwell time of the user at a single base station. However, the first thing we need to make a reasonable assumption. The assumption is that there are no consideration of the time error between base stations. The collected data is calculated according to the calculation formula of the flow characteristics of the prior defined base station coverage area, and the sample data of each base station can be obtained. Some sample data are calculated as shown in Table 3-3. The time unit is minute.

Table 3-3 Part of the calculation results

residence time per person in working hours	percapita residence time in the early morning	percapita residence time at the weekend	daily traffic
36902	78	521	602
36903	144	600	521
36904	95	457	468
36905	69	596	695

In order to eliminate the influence of the difference of the magnitude of each attribute data, this paper carries out the standard deviation treatment on the collected statistical data. This division is based on that the more close to 1, the greater the original data before the standardization processed, the longer the residence time. The processed result is shown in Table 3-4. The results of normalization of variance are plotted by using the Matlab software.

Table 3-4 Deviation Normalization Results

residence time per person in working hours	percapita residence time in the early morning	percapita residence time at the weekend	daily traffic
36902	0.103865	0.856364	0.850539
36903	0.263285	1	0.725732
36904	0.144928	0.74	0.644068
36905	0.082126	0.992772	0.993837

A dendrogram consists of many U-shaped lines that connect data points in a hierarchical tree. The height of each U represents the distance between the two data points being connected. The measure unit is minute. The image is shown in Figure 1.In order to facilitate the study of this paper, we will enlarge the clustering map in order to clearly observe the characteristics. The image is partially enlarged as shown in Figure 2.

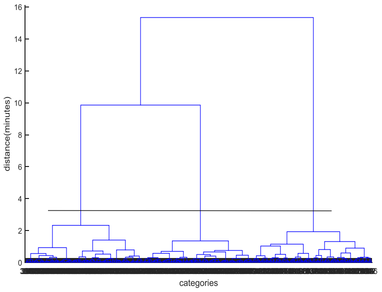

Figure 1 Clustering of Pedigrees

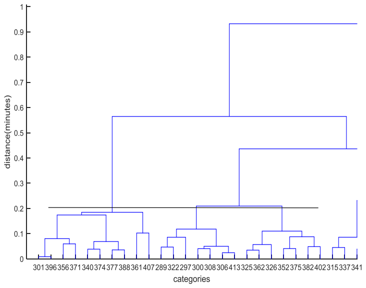

Figure 2 Partially Enlarged

The hierarchical clustering algorithm is implemented in the Matlab. Hierarchical clustering algorithm execution flow chart as follows:

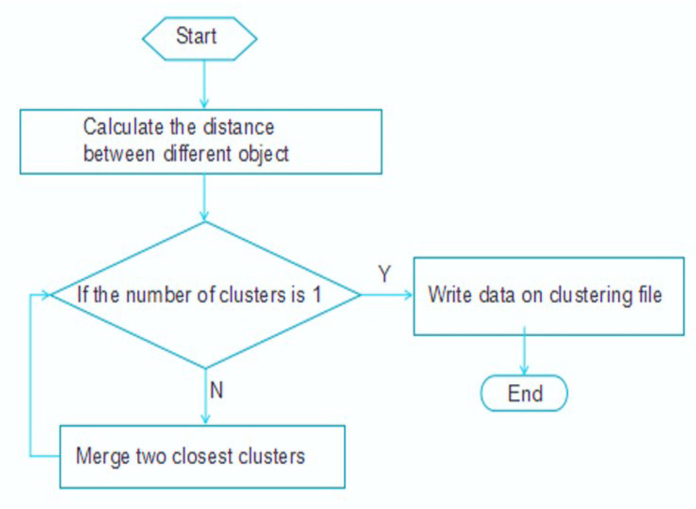
Start
Calculate the distance between different object
If the number of clusters is 1
Y
Write data on clustering file
End
N
Merge two closest clusters

Figure 3 Hierarchical clustering algorithm

We can clearly see that in the Figure1-1 in the U tree, horizontal line in the node is divided into three parts, which is achieved the best clustering effect. So as it can be seen in Figure 1,the standardized data are best classified into three Categories[16,17]. The number of the clusters is set to three. After performing the above four steps, we can draw a cluster diagram with four indexes for the clustering results, as shown in Figure 4, 5, 6.The vertical coordinates represent the residence time after the standardized treatment. There are four abscissa and they are “working hours”, “early morning”, “weekend”, “daily”. They represent the “residence time per person in working hours”, “percapita residence time in the early morning”, “percapita residence time at the weekend” and “daily traffic”.

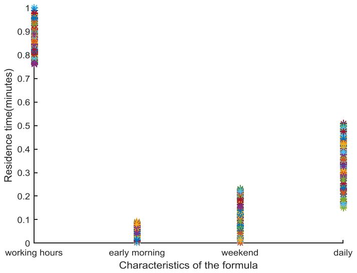

Figure 4 Business Circle Category 1

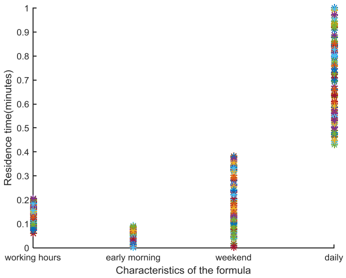

Figure 5 Business Circle Category 2

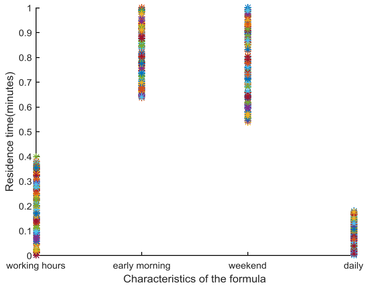

Figure 6 Business Circle Category 3

3.4 Results analysis

In a set of experiments, we have analyzed the results. The contrasts of Figures 4, 5, 6 make it clear that the value of the different business circles is quite different. For Business Circle Category1, we can see that the abscissa is marked as "working hours percapita residence time" and the axis coordinate value is higher, then the user staying time in the coverage area of the first clustering result is longer. At the same time in the morning, the weekend percapita residence time and the average daily traffic volume corresponding to the axis of the coordinate value is relatively low, which is represented the morning and weekend coverage within the user residence time is relatively short in the two time base station. This can be inferred that the region is similar to the white-collar family work area, which is more in line with white-collar workers working area time flow changes. For the Business Circle Category 2, it can be clearly seen from the figure that the first three indicators have a shorter residence time, while the average daily traffic volume has a longer residence time. This can be inferred that the region is similar to the white-collar workers. As it can be seen from the Business Circle Category 3, for all levels, the abscissa is marked as "working hours percapita residence time" and the axis coordinate value is lower. However, the axis coordinate value of “percapita residence time in the early morning" and "percapita residence time at the weekend" are higher than the others. We can get the conclusion that this region are similar to the residential district .

As a result of a huge business area every day, and in the early hours of the morning is generally closed, so Figure 4 more in line with the business district. So this paper can be inferred for the second kinds of cluster base station coverage area belong to the business district. Analysis of Figure 5,we can see in the working day and the average daily traffic volume of the two coordinates of the corresponding coordinate axis value Y is relatively small and the morning and the weekend corresponds to a relatively large axis coordinates, you can see in the observation window during this period of time people stay in the early hours of the morning and a longer time, we can infer a third cluster type base station service area in residential areas.

In summary, based on the analysis of base station log data, the cluster analysis shows that the average daily traffic volume of Business Circle Category1 and Business Circle Category 3 is relatively small, while white-collar workers are concentrated in eating time and commuting time, such areas are not suitable for commercial marketing. For Business Circle Category2, due to its large daily traffic volume, and this part of the base station coverage within the crowd stay relatively long, so it is suitable for commercial marketing.

Conclusion

With the development of mobile communication technology, the use of mobile terminals such as mobile phones are becoming more common, the growth of the mobile phone positioning data for the analysis of the individual activities or understanding urban development provides a powerful data support. In order to increase the business to business circle and carry out the precise marketing business, in this paper, the calculation model of the flow characteristics index is established. By using the hierarchical clustering algorithm, the coverage of the business circle, residential area and work area near the base station is analyzed. The experiments results show that the hierarchical clustering algorithm has good clustering effect and can distinguish the residential area, the work area and the commercial area. It shows the feasibility and effectiveness of the algorithm in data mining. Digging out the high-value information to make recommendations to businesses and to rationalize the promotion of commercial marketing activities. However, this paper does not optimize the distance calculation of the first step data of the hierarchical clustering algorithm, which will inevitably lead to the high time complexity of the algorithm. So my future research work is mainly on the level of clustering algorithm to improve, to further improve the efficiency of the implementation of clustering computing time. Secondly, the hierarchical clustering algorithm is applied to other research fields, extending the use of hierarchical clustering algorithm to study its feasibility in other areas, this is bound to promote the development and application of clustering algorithm.

Acknowledgement

This research was in part supported by grands from the Natural Science Foundation of China (No.61502438), the Hubei Provincial Department of Education Science and Technology Project (No.2014-47).

References

1. T.F. Bao. A Study On Context Recognition and Mining of Mobile User data. University of Science and Technology of China(2012).
2. C. Song, Z. Qu, N. Blumm, et al. Limits of Predictability in Human Mobility. Science, **327**(5968), 1018-1021(2010).
3. X.F. Zhu.Research on Rapid Mining Algorithm for Massive Data.Nanjing University of Posts and Telecommunications(2012).

4. Z.A. Dong, X.Q. Lu. User Behaviour Analyses Based on Baidu Search Logs. Computer Applications and Software, **30**(7), 18-20(2013).
5. Z.G. Z, C.Q. J, X.L. Wang, A.Y. Zhou. etc. Discovering Import Locations from Massiveand Low-Quality Cell Phone Trajectory Data. Journal of Software, **27**(7), 1-14(2016).
6. Y.X. Kong, C.Q.Jin, X.L. Wang. etc. Population Flow Analysis Based on Cellphone Trajectory Data. Journal of Computer Applications, **36** (1), 44-51(2016).
7. J. Chen, B. Hu, X.Q. Z, Y. Yang.Personal Profile Mining Based on Mobile Phone Location Data. Geomatics and Information Science of Wuhan University, **39**(6), 734-738(2014).
8. S.F. Gong, W.L. Chen, P.T. Jia.Surver on Algorithms of Community Detection. Application Research of Computers, **30**(11), 3216-3227(2013).
9. T. Bao, Z.G. Zhang, C.Q. Jin.An Urban Population Flow Analysis System Based on Mobile Big Data[J]. Journal of East China Normal University(Nature Science), **9**(5), 162-170(2015).
10. Z.C. An. Mining User Mobility Behavior Based on Base Station. Harbin Institute of Technology(2011).
11. N. X, L. Y, J.X. Hu. Identifying Home-Work LocationsFromShort-Term,Large-Scale, and Regularly Sampled Mobile Phone Tracking Data. Geomatics and Information Science of Wuhan University,**39**(6), 750-755(2014).
12. J. Shen, L. Z, J.W. Yang, R. Li. Archical Clustering Algorithm Based onPartion. Computer Engineering and Applications, **43**(31), 175-177(2007).
13. Y.Miu. The Analysis of Data Based on TheHierarchical Clustering.Anhui University (2013).
14. H. Ying, K.J. Xia, Data preprocessing method based on user characteristic of interests for web log mining. Proceedings-2014 4th International Conference on Instrumentation and Measurement, Computer, Communication and Control, 867-872(2014).
15. Montoliu R, GaticaPerez D. Discovering human places of interest from multimodalmobile phone data.Proceedings of the 9th International Conference on Mobile andUbiquitous Multimedia. ACM, **12**(2010).
16. Blumenstock J E. Using mobile phone data to measure the ties betweennations. Proceedings of the 2011IConference. ACM, 195-202(2011).
17. Park M, Lee T. Understanding science and technology information users through Iransaction loganalysis. LibraryHi Tech, **31**(1), 123-140(2013).