Detecting Urban Functional Area Based on Clustering on Base Station Data

Yuan Cheng[1], Yahong Wang[2], and Ronghua Chi[3]
{changuang7@sina.com[1], wyh811619@126.com[2], chironghua@126.com[3]}

College of Computer Science and Technology, Harbin University of Science and Technology[1],
School of Applied Science and Civil Engineering, Beijing Institue of Technology, Zhuhai[2]
College of Computer and Information Engineering, Heilongjiang University of Science and Technology[3]

Abstract. Detecting urban functional area can contribute to rational planning of urban functions, which is conducive to urban construction. Because the call data correlates with the users' locations and their behavior characteristics, we use the uncertain clustering method to find the aggregate characteristics of the users' calling behaviors, based on the data produced by base station providing the communication services, and then analyze the related land functions. Finally, we will verify the effectiveness of the proposed method through experimental results.

Keywords: Urban functional area, uncertain clustering, base station data, calling behaviors.

1 Introduction

Big data technology and machine learning methods have provided strong technical support for the development of smart city, which mainly focuses on mining knowledge from various types of data (such as taxi trajectory, call detail records, etc.) to assist in city construction. Detecting urban functional areas from user call detail records or travel behaviors contributes to urban planning. The obtained results are more effective, accurate, and more conducive to discovering the relationship between land functions and human activities, compared with traditional methods, such as field investigations or remote sensing technologies. That is because call record data or trajectory data could reflect human activities in a certain place. Some existing studies choose to judge the function of user calling locations through

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2296733

discovering the characteristics of the users' calling behaviors based on the users' call record data[1-4]. While some studies choose to judge the function of the human movement area by analyzing the characteristics of the users' moving behaviors based on the their movement trajectory data [5-8]. However, because that the data used for analysis are mostly related to personal behavior, it may involve user privacy protection issues. The data mining methods used to discover human behavior characteristics mainly include two types, one is classification method [9-11], and the other is clustering method [12-14]. Discovering urban functional areas is mainly to explore the relationship between human behavior characteristics and regional functions, so it is not necessary to conduct analyzing based on detailed data representing individual behaviors. The data can also be appropriate as long as it could be used to analyze the aggregate characteristics of human behaviors related to land functions. Apart from that, the classification method is not suitable for the case that there is no labeled data.

Therefore, for the purpose of privacy protection, we will propose a method to analyze the aggregation characteristics of the human calling behaviors based on the statistical data about the number of persons served by the base stations in the city, and then to explore the related urban functional areas.

2 Data Processing

We will use the data collected from the base stations of Harbin providing the services of communication. Each data record represents the number of service persons monitored by each base station in the city at the collection time point. Its specific information include the geographic location of the base station, the collection time, and the number of people served, and the base station ID. There are about 3,096 base stations and about 130 million records in the dataset. The main attributes of the original data and some data examples are shown in Table 1.

Table 1. Examples of base station data.

Number of users in DCH	Number of users in DSCH	Collection time	The wireless network controller	Affiliated region	Longitude	Latitude
26	18	2013/2/4 1:00	HRBRNC04	Nangang	126.6833	45.75097
24	20	2013/2/4 1:01	HRBRNC04	Nangang	126.6833	45.75097
…	…	…	…	…	…	…
29	20	2013/2/4 2:00	HRBRNC04	Nangang	126.6833	45.75097
…	…	…	…	…	…	…

However, when using the data in Table 1, there may exist some objective conditions that will result in some uncertainties in the data set: the users are likely to move from one base station to another base station's coverage area near the collection time point, and it is also possible that the number of serving persons is zero in the areas with fewer people at a collection time point. The simple data preprocessing techniques such as dimensionality reduction and denoising, may ignore the actual significance of these data. Although the data is uncertain at a certain collection time point, in a coarser-grained data set, such as the sample data produced by a certain base station within one hour, the data are more likely to have the significance of reflecting the service mode of the base station. That can be explained by the law of large numbers: the uncertainty displayed by the average behavior of multiple individuals in a system will gradually decrease as the total number of individuals continues to increase.

Therefore, we will make our analysis with coarse-grained data, which is constructed by the original data with form of Table 1 in hours. Then each data record represents the information about the number of users served by the base station in one day, and is constituted by the 24-hour data. Its form is shown in Table 2. Each attribute is composed of samples collected from 1-hour data in the original dataset. That is, the data of users served by one base station which is observed at each collection time (i.e. every minute) in the original data set is used as a sample in one attribute of the coarse-grained data set, for the purpose of reflecting the distribution of users in the coverage area of each base station within each hour.

Table 2. The coarse-grained data after processing.

Number of users in DCH from 0:00 to 0:59	Number of users in DCH from 1:00 to 1:59	…	Number of users in DCH from 23:00 to 23:59	Collecti-ng date	The wireless network controller	Affili-ated region	Longitude	Latitude
26,24, …	14,10,…	…	37,40, …	2013/2/4	HRBRNC04	Nan-gang	126.6833	45.75097
…	…	…	…	…	…	…	…	…
19,20, …	21,25, …	…	22,17,…	2013/2/8	HRBRNC04	Nan-gang	126.6093	45.7203
…	…	…		…	…	…	…	…

The data reflect the distribution of users served by each base station in each time period. we will propose an uncertain clustering method on this data set to analyze the aggregated features of calling behaviors, and then detect the functions of urban areas based on the calling features.

3 Uncertain Clustering on Base Station Statistical Data

We can treat each data described in Table 2 as an uncertain object, and it represents the information about one base station providing services in one day. The reason to call it as an uncertain object is that it consists of some samples which are collected to represent the information about one base station providing services in each hour. Therefore, in order to analyze the uncertain objects efficiently, firstly we use the method proposed in the previous research result [15] to build an uncertain data model, by which we can describe the distribution of the number of users covered by one base station in each time period. Then we propose an clustering method to aggregate data represented by the uncertain data model.

The data model is constructed for each base station data after processing according to the method proposed in [15] is shown as equation (1).

$$\begin{aligned} G(x_j) &\simeq \sum_{sim(x_j - c_k) \geqslant h_\rho} \sum_{|\alpha| \leqslant p} C_\alpha^k \mathrm{e}^{-\frac{\| x_j - c_k \|^2}{h^2}} \left(\frac{x_j - c_k}{h} \right)^\alpha \\ &= \sum_{sim(x_j - c_k) \geqslant h_\rho} \sum_{(|\alpha_1| + \ldots + |\alpha_d|) \leqslant p} C_\alpha^k \mathrm{e}^{-\frac{\| x_j - c_k \|^2}{h^2}} \cdot \left(\left(\frac{x_j - c_k}{h} \right)_1^{\alpha_1}, \ldots, \left(\frac{x_j - c_k}{h} \right)_d^{\alpha_d} \right) \end{aligned} \tag{1}$$

where $C_\alpha^k = \frac{2^{|\alpha|}}{\alpha!} \sum_{s_i \in S_k} q_i \mathrm{e}^{-\frac{\| s_i - c_k \|^2}{h^2}} \left(\frac{s_i - c_k}{h} \right)^\alpha$, and one uncertain data record (i.e. one object) in the dataset $x_j, 1 \leqslant j \leqslant M$, the sample set $S_j = \{s_{j1}, \cdots, s_{jn}\}, 1 \leqslant i \leqslant M$ which represent the uncertain object x_j, bandwidth h, the similarity threshold h_ρ, and the degree threshold of polynomial retention terms ρ are the parameters of the data model construction method. Their specific explainations can be found in [15]. The dataset described as Table 2 could reduce the uncertainties existing probably in the original statistical data, and this modelling method could ensure the high computational efficiency of model construction.

A clustering mehtod is needed for mining the aggregating features of uncertain data,. Aiming at the characteristics of uncertain data, the basic idea of traditional clustering algorithms can be extended to the field of uncertainty. Considering that K-Means is suitable for processing numerical attributes and has the characteristics of linear time complexity for processing large data sets, we will use the idea of K-Means to cluster uncertain data objects. When facing uncertain data objects, we need to process the uncertainties in data objects and in the clusters involved in the algorithm.

The uncertain data can be represented as equation (1). And the model describing the distribution of mean value of cluster i consisting of n_i uncertain objects can be represented as equation (2), where K is the number of clusters. That is, similar to the uncertain object, the cluster mean is obtained from the distribution characteristics that represent the uncertainty of the objects in the cluster, and can be obtained by calculating the average of the probability density functions of all the objects in the cluster.

$$pdf\,(mean_i) = \tfrac{1}{n_i}\sum\nolimits_{j=1}^{n_i} G(x_j), 1 \leqslant i \leqslant K \tag{2}$$

Then we can define the similarity between uncertain objects represented by polynomials as the form of equation (1). If the coefficients of the two polynomials are expressed as vectors $\boldsymbol{v}_1 = (\boldsymbol{v}_{11}, \boldsymbol{v}_{12}, \cdots, \boldsymbol{v}_{1m})$, and $\boldsymbol{v}_2 = (\boldsymbol{v}_{21}, \boldsymbol{v}_{22}, \cdots, \boldsymbol{v}_{2m})$ respectively, the similarity is defined in equation (3), i.e. it can be obtained by calculating the cosine similarity between two objects, whose data model are $g(x_1)$ and $g(x_2)$, where m is the larger number of terms in the polynomials of uncertain data model. Once we describe the cluster mean as equation (2), we can also use equation (3) to calculate the similarity between an uncertain object and a cluster.

$$p_s = sim(g(x_1), g(x_2)) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{\| \boldsymbol{v}_1 \| \ \| \boldsymbol{v}_2 \|} \tag{3}$$

Based on the data model of the uncertain object, the similarity measurement between uncertain objects, and the definition of the cluster mean, we propose the method of clustering on uncertain data based on K-Means as follows.

```
IUK-Means(D, $G(x_i)$, $K$)
Input: uncertain dataset $D = \{x_1, x_2, \ldots, x_M\}$,
         uncertain data model $G(x_i), 1 \leqslant i \leqslant M$,
         number of clusters $K$;
Output: clustering result $C = \{C_1, \cdots, C_K\}$;

1) select $K$ objects randomly from $D = \{x_1, x_2, \ldots, x_M\}$ as the initial cluster center;
2) repeat
3)   for $i = 1$ to $M$ do
```

4) Calculate the similaritys between object x_i and the mean of each cluster, that is, $sim = \{sim_1,\ sim_2, \cdots, sim_K\}$ based on equation (3);

5) Assign the object x_i to the most similar cluster (i.e. $sim_j = \max(sim_1,\ sim_2, \cdots, sim_K)$) to the cluster C_j;

6) **end** for

7) Update the data model of K clusters to describe their distribution features, $pdf(mean_j)$, $(1 \leqslant j \leqslant K)$ according to equation (2);

8) Calculate the objective function $E = \sum_{j=1}^{K} \sum_{o \in C_j} sim(o, mean_j)^2$;

9) **until** the objective function E reaches convergence.

The time complexity of this method based on the idea of K-Means for clustering uncertain objects is $O(tKdM)$, where t is the number of iterations, and d is the dimension of the uncertain object.

4 Experimental Results

We apply the above method on the data as shown in Table 2. 18 clusters are generated, and different clusters have different features. The features are shown in **Figure 1**. We can summarize the functional areas of these clusters in Table 3 based on the features of the calling behaviors.

Table 3. The clusters and their functional areas.

ID	Cluster	Functional area
1	Cluster 1	Mixed area
2	Cluster 3	Enterprise area
3	Cluster 7 Cluster 9	Tourist area
4	Cluster 10	Residential area
5	Cluster 15	Tourist area
6	Cluster 18	Mixed area

7	Cluster 2, 5, 6,8,11,13,14, 16, 17	Others
8	Cluster 4 Cluster 12	Outliers

In the 18 clusters, the clusters in the 7th group of Table 3 seem to have similar features according to Figure 1. It seems that there are only a few people at a certain time point, and there are no people to be served at the other time. That may be areas with fewer people, which are likely to be related to local job characteristics. People covered by stations in Cluster 4 and 12 are approximate zero, so we cannot infer their corresponding functions. Stations in Cluster1 and 18 serve people with similar behavior characteristics. The difference is that the number of people covered by base stations starts to decrease after 23 o'clock in Cluster 1, while after 16 o'clock in Cluster 18. Because it remains a relatively stable number of people all over the day, it can be infered as a mixure of office area and residential area. The person number in Cluster 3 starts to increase from 8 a.m., reaches a peak during 10 a.m. and 11 a.m., and starts to decrease significantly from 4 p.m. to 5 p.m.. Such areas are generally enterprise or institution areas with obvious rules. Cluster 7 and Cluster 9 are characterized by an increase starting from 8 a.m. and decrease from 8 p.m.. Their peaks appear at 3 p.m.. and no one appears in other time periods. Similarly, The person number of Cluster 15 starts to increase from 8 a.m. with the peak appearing at 8 p.m., and basically approximates to zero after 23 o'clock. The characteristics indicate that these base stations in the three clusters are in tourist areas, and the best time to visit the scenic spots are at the time near the peak. The number of people starts to increase after 15 o'clock and remains stable 24 o'clock, which are the features like residential areas, that is, the people covered by the base stations start to increase since the time of after work.

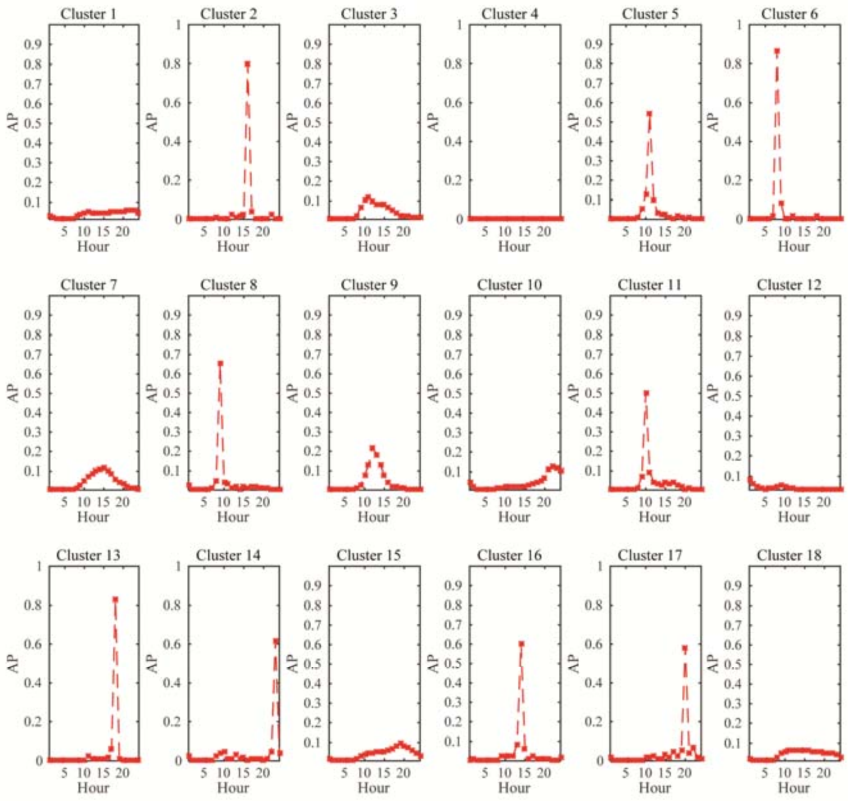

Fig. 1. Features of clusters generated on data described as Table 2.

Obviously, we could detect the urban functional areas according to the behavior characteristics of base stations serving people without user privacy data. If we can gather information provided by more base stations in a city, we could obtain more detailed division of functional areas.

Conclusion

In order to solve the problems about privacy leakage when analyzing the functions of the urban areas based on the detailed data of individual behaviors, this study uses the statistical data from Harbin base stations, and proposes a method to discover the land functions based on uncertain clustering. This study uses the data monitored at each gathering time as sample points to form coarser-grained data sets, builds uncertain data model and performs uncertain clustering on this basis. The mehtod could not only reduce the impact of the possible uncertainties existing in original data, but also detect land functions through the related aggreated features of calling behaviors based on clustering results. The experimental results indicate the effectiveness of the proposed method .

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References

[1]Toole J L, Ulm M, González M C, et al.: Inferring land use from mobile phone activity. Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, pp. 1-8 (2012)

[2]Pei T, Sobolevsky S, Ratti C, et al.: A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science,Vol. 28(No. 9), pp. 1988-2007(2014)

[3]Calabrese F, Ferrari L, Blondel V D. Urban sensing using mobile phone network data: a survey of research. Acm computing surveys (csur),Vol. 47(No. 2), pp. 1-20(2014)

[4]Louail T, Lenormand M, Ros O G C, et al.: From mobile phone data to the spatial structure of cities. Scientific reports,Vol. 4, pp. 5276(2014)

[5]Zhang X, Li W, Zhang F, et al.: Identifying Urban Functional Zones Using Public Bicycle Rental Records and Point-of-Interest Data. ISPRS International Journal of Geo-Information,Vol. 7(No. 12), p. 459(2018)

[6]Caceres N, Benitez F G.: Supervised land use inference from mobility patterns. Journal of Advanced Transportation,Vol. 2018, p. 8710402(2018)

[7]Liu Y, Wang F, Xiao Y, et al.: Urban land uses and traffic ‘source-sink areas’: Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, Vol. 106(No. 1), pp. 73-87(2012)

[8]Xia T, Li Y.: Revealing Urban Dynamics by Learning Online and Offline Behaviours Together. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,Vol. 3(No. 1),Article 30(2019)

[9]Wu Y, Qiao Y, Yang J, et al.: Urban Functional Area Division Based on Cell Tower Classification. IEEE Access,Vol. 7, pp.171503-171514(2019)

[10]Ge P, He J, Zhang S, et al.: An Integrated Framework Combining Multiple Human Activity Features for Land Use Classification. ISPRS International Journal of Geo-Information,Vol. 8(No. 2), pp. 90(2019)

[11]Tu W, Cao J, Yue Y, et al.: Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. International Journal of Geographical Information Science,Vol. 31(No. 12), pp. 2331-2358(2017)

[12]Tang L, Gao J, Ren C, et al.: Detecting and evaluating urban clusters with spatiotemporal big data. Sensors,Vol. 19(No. 3), p. 461(2019)

[13]Manley E, Dennett A.: New Forms of Data for Understanding Urban Activity in Developing Countries. Applied Spatial Analysis and Policy,Vol. 12(No. 1), pp. 45-70(2019)

[14]Yao Y, Li X, Liu X, et al.: Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. International Journal of Geographical Information Science, Vol. 31(No. 4), pp. 825-848 (2017)

[15]Yuan Cheng, Ronghua Chi, Suxia Zhu, et al.: An Uncertain Data Model Construction Method Based on Nonparametric Estimation. Proceedings of 2016 IEEE International Conference on Electronic Information and Communication Technology, 2016, Harbin. USA: Institute of Electrical and Electronics Engineers Inc,pp. 384-389 (2016)