

# Important Location Identification and Personal Location Inference Based on Mobile Subscriber Location Data Preparation of Camera-Ready Contributions to SCITEPRESS Proceedings

YANG Zhen , WANG Hong-jun

College of Electronic Countermeasure, National University of Defense Technology, Hefei , China

**Abstract:** As an emerging spatial trajectory data, mobile terminal location data can be widely used to analyze the behavior characteristics and interests of individuals or groups in smart cities, transportation planning and other civil fields. It can also be used to track suspects in anti-terrorism security and public opinion management. Aiming at the problem that it is difficult to determine suitable input parameters of clustering caused by different subscriber location data size and distribution difference, an improved density peak clustering algorithm is proposed and the performance of the improved algorithm is verified on the UCI data set. Firstly the important location is identified by the proposed algorithm, and the personal location is further inferred by the algorithm based on the subscriber's schedule and maximum cluster. Then, the algorithm adopts Google's inverse geocoding technology to obtain the semantic names corresponding to the coordinate points, and introduces the natural language processing technology to achieve word frequency statistics and keyword extraction. The simulation results based on the Geolife data set show that the algorithm is feasible for identifying important locations and inferring personal locations.

## 1 INTRODUCTION

In recent years, with the rapid development of mobile communication technologies and the increasingly powerful functions of smart mobile terminal networks, smart terminal devices such as mobile phones and tablet computers have gradually surpassed personal computers and become the most widely used information devices for people. At the same time, the rapid development of global navigation and positioning systems has provided accurate and real-time location information for smart mobile terminals. Location Based Service (LBS) has become one of the most popular terminal information services. A large number of related researches have also proved that the important information mining value is implied in massive location data.

Important location identification refers to the region where the user has a long stay and a high frequency of visits from the historical location data of the target user[1]. The personal location inference is based on the identification of important locations, further analysis of the extracted areas, thus inferring the user's home address, workplace or entertainment, etc, to discover the user's activity patterns, establish their behavior patterns[2], and speculate user's hobbies, even health and income levels, etc. The user information obtained through mining can not only recommend personalized interest points for

users at the commercial level, but also can track suspicious targets at the security level.

The key point of important location identification technology is to definite the identification threshold. If the threshold is set too low, it will cause many insignificant points to be mixed in the recognition results. On the contrary, some important locations will be filtered out by mistake.

Yang et al. first used the gradient threshold to process the noise point, then used the DBSCAN algorithm to cluster the location points and output all kinds of clusters as important locations[3]. Montoliu et al. first extracted the stay points using a time-based clustering algorithm, then extracted the stay regions using grid-based clustering, and output them as important locations[4].

The key to personal locations inference technology is to find the inference principle that is suitable for most users. Hoh et al. proposed a home address recognition algorithm based on the time period of work and rest[5]. Krumm et al. collected the car trajectory of 172 volunteers at a frequency of 6 seconds in two weeks by installing a GPS receiver on the car, and judged the last arrived position on the day as the user's home address[6]. Zang et al.'s research shows that the location of the base station where the user's mobile phone is located has the highest frequency of two base stations often represents the user's place of residence and work place[7].

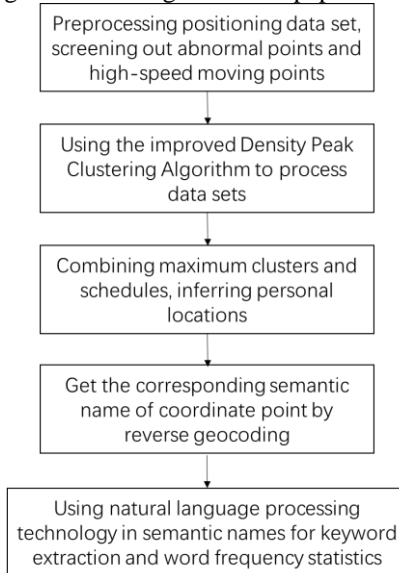
In general, the current method focuses on the mining of the spatial level of location data, but not much

attention is paid to the information contained in the semantic level.

## 2 ALGORITHM PRINCIPLE

### 2.1 General Program

The algorithm is designed in this paper as follows:



### 2.2 Density Peak Clustering Algorithm and Its Improvement

Clustering refers to the unsupervised classification according to a certain characteristic of the research object, ie, the minimal similarity among different clusters and the greatest similarity in the same cluster can be obtained without the pre-set the classification criteria according to the similarity difference of the research object.

The density-based clustering algorithm clusters datasets according to the density of data points. The main idea is to find high-density regions separated by low-density regions. The advantage is that clustering is effective for irregular clusters. Therefore, it is widely used in cluster analysis of spatial locations.

Being a classic density-based clustering algorithm, DBSCAN algorithm is widely used. The algorithm is very sensitive to the input parameters and the effect of clustering on non-uniform density data sets is not too ideal since it needs to set two globally unique input parameters artificially. However, in the massive location data, the location points generated by different subscribers have great differences in scale and distribution shape, which undoubtedly puts forward higher requirements for the flexibility of parameter setting of the clustering algorithm.

Alex Rodriguez and Alessandro Latio proposed a new density-based clustering algorithm called Density Peaks Clustering Algorithm (DPCA)[8]. The algorithm is simpler than the DBSCAN algorithm because only one parameter is needed to input and iterations are not required. Cluster analysis can be performed on various shapes of point clusters. However, it is necessary to

manually select cluster centers through decision maps. This not only increases the redundancy of the algorithm, but also exists subjectively hidden dangers.

#### 2.2.1 Density Peak Clustering Algorithm

The only input parameter of the Density Peak Clustering Algorithm is the cut-off distance  $d_c$ . This paper adopts the percentage to represent the cut-off distance  $d_c$ , arrange the distances between all two data points in ascending order, and take the distance value corresponding to a certain percentage in the distance set.

In this model, for a data set  $S=\{x_i\}_{i=1}^N$ , to be clustered, there are two quantities that need to be calculated. One is the local density of each data point  $\rho_i$ , and the other is the minimum distance between each data point and the higher density point  $\delta_i$ .

The formula for the local density is as follows:

$$\rho_i = \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \quad (1)$$

$\delta_i$  is the minimum distance from data point  $i$  to data point  $j$  with a higher local density, defined as:

$$\min_{\rho_j > \rho_i} (d_{ij}) \quad (2)$$

For the data point  $k$  with a global maximum density,  $\delta_k$  is defined as the maximum distance between the data point and the remaining data points.

At this point, for each data point in the data set  $S$ , its local density  $\rho_i$  and the minimum distance  $\delta_i$  can be calculated. The density peak clustering algorithm regards the data points that have large  $\rho_i$  and  $\delta_i$  at the same time as the cluster centers, and the remaining data points are allocated to clusters that are closest to the nearest cluster centers. As shown in Figure 1, the horizontal axis represents the local density of each point, the vertical axis represents the minimum distance of each point, and the color points are the points selected as the cluster centers.

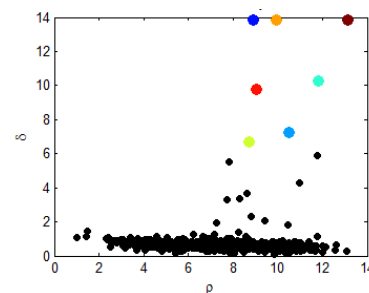


Figure 1: Decision diagram

The density peak clustering algorithm mainly has two shortages. On the one hand, there is no reliable reference for the selection of its input parameter  $d_c$ . The author of this algorithm Rodriguez gave the recommendation to select the top 1% to 2% of the cut-off distance after the distance between all data points in ascending order. However, for a data set with a disproportionately different size and distribution, this general method of parameter selection will often lead to poor results. On the

other hand, the density peak clustering algorithm relies on the distribution of points on the decision graph to select the cluster center, which is too subjective.

Use 15-point type for the title, aligned to the center, linespace exactly at 17-point with a bold font style and initial letters capitalized. No formulas or special characters of any form or language are allowed in the title.

Words like “is”, “or”, “then”, etc. should not be capitalized unless they are the first word of the title.

### 2.2.2 The Improved Density Peak Clustering Algorithm

The ADPC algorithm is an improved algorithm in this paper. Its essence is to achieve the data set clustering through the adaptation of the cut-off distance and the automatic selection of the clustering centers.

Rodriguez gave an idea for reference[8], that is, consider the product of  $\rho_i$  and  $\delta_i$  :

$$\gamma_i = \rho_i \delta_i \quad (3)$$

First,  $\rho_i$  and  $\delta_i$  are normalized, then  $\gamma_i$  is obtained for each data point, and arrange them in descending order. The larger the  $\gamma_i$ , the more likely it is to be a cluster center.

To visually show the relationship between  $d_c$  and  $\gamma_i$ , formula (3) can also be expressed as:

$$\gamma_i = \delta_i \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \quad (4)$$

Finding the minimum value of the information entropy based on  $\gamma_i$  to achieve the adaptive truncation distance and obtaining the maximum value of the slope change trend to achieve automatic selection of cluster centers is the focus of this paper.

#### 2.2.2.1 The Adaptation of Cut-off Distance

This paper proposes a cut-off distance adaptive method based on the minimization of information entropy. In information theory, entropy is used as a measure of system uncertainty. The greater the entropy, the stronger the uncertainty of the system. The formula for calculating information entropy is as follows:

$$H = -\sum_{i=1}^n p_i \log p_i \quad (5)$$

Similarly, apply the definition of entropy to this algorithm, Replace the probability value in the information entropy formula with  $\gamma$ , that is, the larger the  $\gamma$ , the more likely it is that the corresponding data point is the cluster center. If the points in the dataset have the same  $\gamma$ , the uncertainty of the system is the greatest, and the difficulty of determining the cluster center is also highest at this time. On the other hand, if the entropy value is the minimum, the distribution difference of  $\gamma$  is

the most obvious, and it is the easiest to determine the cluster center.

Combine equations (4) and (5), this paper gives the relationship between cut-off distance  $d_c$  and entropy  $H$ :

$$H = -\sum_{i=1}^n \left( \delta_i \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \right) \log \left( \delta_i \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \right) \quad (6)$$

Let the cut-off distance  $d_c$  gradually increase from zero, find  $d_c$  that makes the entropy value  $H$  the minimum, and use it as the optimal cut-off distance for the next clustering, so that the cut-off distance is self-adaptive.

Take the data set of [8] as an example and get the result shown in Figure 2. The horizontal axis indicates the cut-off distance, and the vertical axis indicates the entropy value. It can be seen from the figure that when the entropy value reaches the minimum, the cut-off distance is 1.5%. It can be seen that this method can achieve self-adaptation of the truncation distance.

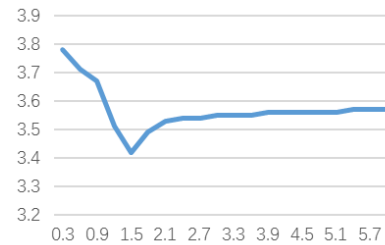


Figure 2: The effect of cut-off distance on entropy value

#### 2.2.2.2 Automatic Selection of Cluster Centers

Take the 2000 random sample points given in [8] as an example, select the first 1.5%  $d_c$ , calculate the first 40  $\gamma_i$ , and then use MATLAB to generate a decision map based on the results. The results are shown in Figure 3. The first 5 points are the actual cluster centers, and there is a steep downward trend from the cluster center to the first non-cluster center. Therefore, this paper uses the trend of slope change to automatically select the cluster center and define this trend as  $tend_i$ :

$$tend_i = (i-1) \frac{\gamma_{i-1} - \gamma_i}{\gamma_i - \gamma_{i+1}} \quad (7)$$

Considering that there may be a tendency similar to the slope change from point 1 to point 2 in Figure 3, the weight  $i-1$  is introduced for  $tend_i$ . However, as the number of data points increases, the weight  $i-1$  may cause  $tend_i$  to increase rapidly, resulting in erroneous judgments.

In this regard, the paper uses the average value of all  $\gamma_i$  in the ranking graph as the threshold  $\theta$  to solve the problem that the weight will cause the rapid increase of  $tend_i$ . First, the data points whose  $\gamma_i$  is larger than the threshold  $\theta$  are selected from 40 data points, and  $tend$  of the selected data point is calculated. As shown in Figure 4, after the screening, there are still seven data points remaining. It can be seen that the  $tend$  at the sixth point is the largest, so the first five points are selected as potential

cluster centers. The density peak clustering algorithm can well classify the data points with very small  $\rho_i$  and large  $\delta_i$  into outliers. However, relying solely on the product of  $\rho_i$  and  $\delta_i$  to select cluster centers may result in misselection of cluster centers: For the point with large  $\rho_i$  and small  $\delta_i$ , that is, two high density points within the same cluster are likely to be mistakenly selected as clustering centers and the cluster is divided. Therefore, the optimal  $d_c$  obtained by formula (6) is set as the distance threshold of the potential cluster center, and the algorithm is prevented from erroneously judging the point with large  $\rho_i$  and small  $\delta_i$  as the cluster center.

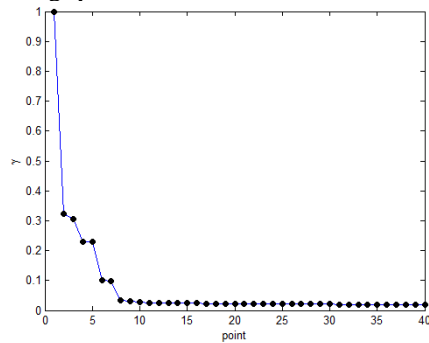


Figure 3:  $\gamma_i$  sort diagram

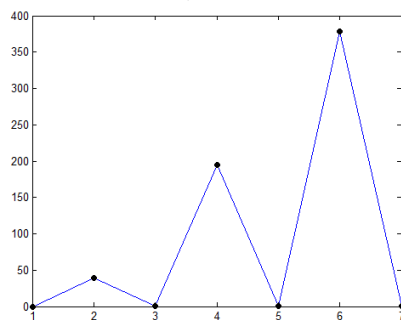


Figure 4: Slope change trend graph

### 2.2.3 The Flow of Improve Algorithm

The concrete steps of ADPC are as follows:

Step 1. Calculate the product  $\gamma_i$  of normalized  $\rho_i$  and  $\delta_i$  for each data point.

Step 2. Find  $d_c$  that minimizes the entropy.

Step 3. Sort  $\gamma_i$  of each data point in descending order, and select the appropriate number of points according to the total number of points in the data set and generate sorting charts in descending order.

Step 4. Calculate the average value  $\theta$  of  $\gamma$  in the  $\gamma_i$  sort diagram and use  $\theta$  as the threshold to filter out data points with  $\gamma_i$  greater than  $\theta$ .

Step 5. According to the slope change trend formula, the *tend* of the filtered data points is obtained.

Step 6. The  $i-1$  data points before the data point  $i$  with the largest *tend* are regarded as potential cluster centers.

Step 7. First, use the first data point as the actual cluster center to determine whether the distance between the second data point and its distance is less than the distance threshold  $d_c$ . If it is less than, then the second data point is treated as a non-clustering center; otherwise, use the second data point as the actual clustering center.

Step 8. By analogy, determine whether the distance between the  $k$ th potential cluster center and all actual cluster centers are greater than the distance threshold, and then treat it as the actual cluster center or non-cluster center according to the judgment result.

Step 9. Finally, according to all the selected actual cluster centers, cluster the remaining data points.

## 3 SIMULATION EXPERIMENTS AND ANALYSIS

### 3.1 Comparison of Algorithm Clustering Performance

In order to verify the clustering effect of the ADPC algorithm, compare the accuracy, normalized mutual information (NMI) and F-measure values of DBSCAN algorithm, density peak clustering algorithm and ADPC algorithm on the UCI dataset. The experimental environment is Windows 10 64-bit operating system, Intel Core i7-6700HQ @2.60GHz CPU, 8G memory, use MATLAB2014a for simulation experiments, and data set information is shown in Table 1.

Table 1: UCI dataset used in the experiment

Data set	Category	Sample	Dimension
Iris	3	150	4
Aggregation	7	788	2
Waveform	3	500	21
Wine	3	178	13

The experimental results are shown in Figure 5~7.

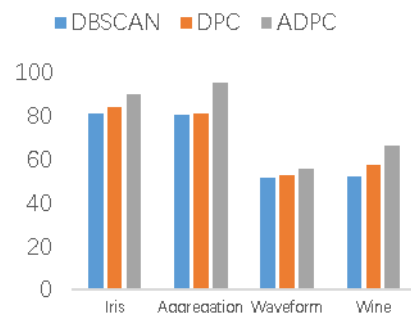


Figure 5: The accuracy of each algorithm

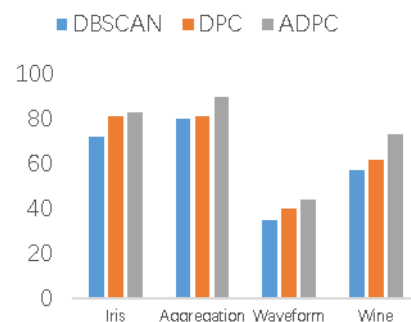
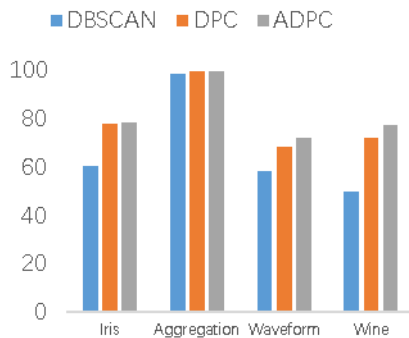


Figure 6: F-measure value of each algorithm



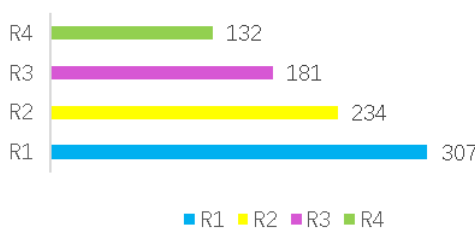
**Figure 7:** The normalized mutual information value of each algorithm

From Fig. 5 to Fig. 7, it can be seen that the trend of ADPC>DPC>>DBSCAN on the accuracy rate, F-measure value, and normalized mutual information of the algorithm as a whole indicates to some extent the ADPC algorithm's superior performance on clustering. However, the performance of the ADPC algorithm on the high-dimensional datasets such as Waveform and Wine is still not ideal. For two-dimensional dataset Aggregation, the accuracy, F-measure value, and normalized mutual information of the ADPC algorithm have reached more than 90%, and the performance is good. Therefore, the ADPC algorithm is used for two-dimensional GPS location points.

### 3.1 Important Locations Identification and Personal Locations Inference

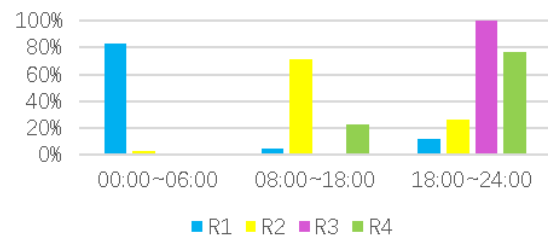
This paper uses the seven-day positioning point of the user 159 in the Geolife dataset to test the identification of important locations. It is labeled with Google Maps. The initial number of points is 1293. After the speed pruning, the nonsense high-speed points are filtered out. After the outliers, the number of points clustered by the ADPC algorithm is 854, and the clustering rate is 66%.

Next, count the number of cluster points for each cluster, as shown in Figure 8.



**Figure 8:** The location point after pruning and clustering

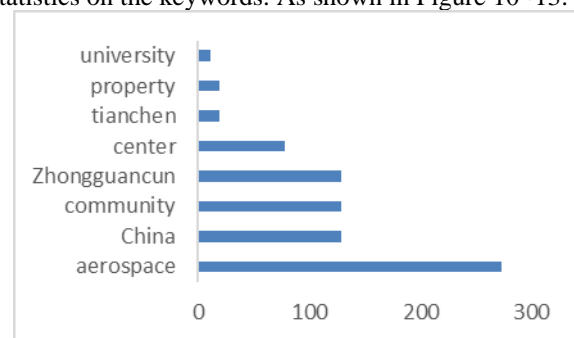
As can be seen from Figure 10, the first two regions with the largest number of clustering points are R1 and R2. According to the conclusion[7], these two regions are usually the places of residence and workplaces. Next, the method of work-rest time period is used to count the proportion of time in each area, as shown in Figure 9.



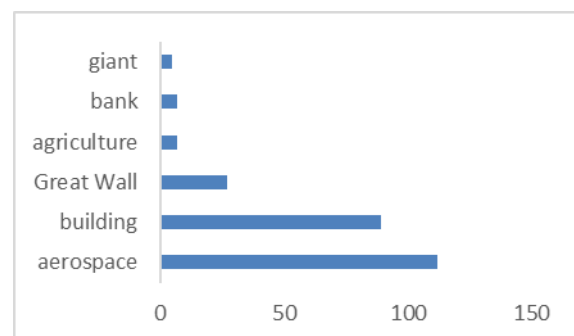
**Figure 9:** The proportion of time in each region

As can be seen from the figure, during the period from 00:00 to 06:00, the proportion of R1's time is much higher than that of R2. Therefore, R1 is marked as "Home Address". In the period from 08:00 to 18:00, the proportion of R2's time is much higher than that of R1. Therefore, R2 is marked as the "work place". The period of 18:00~24:00 usually is personal leisure time, nearly 30% of the R2 area is located in this period, so the user is supposed to work overtime occasionally. All location points in the R3 area are in this period, that is, the user only moves in the R3 area during this period of time, and R4 also has nearly 80% of the location points in this period, so it is inferred that R3 and R4 are the user's entertainment places or other places that are often visited at leisure. Even if the above method has identified the user's home address, work place and the entertainment places, the information of particle size is still large, not enough to reflect the user's personalized features.

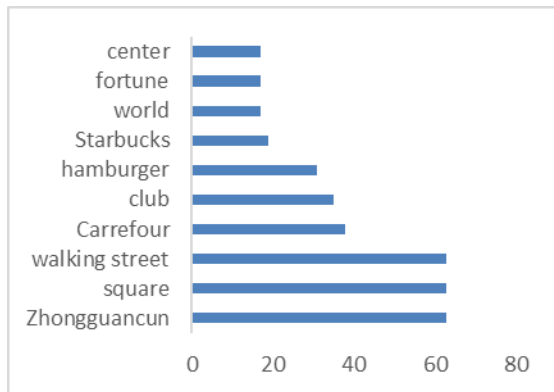
Therefore, after the semantic tagging in each region is given, the coordinates of all the clustering points are entered through the Google reverse geocoding service, and the name information of each clustering point is obtained. Then, the open-source Chinese word frequency analysis software was used to analyze the name information of cluster points in R1~R4, find out the keywords in the name, and perform word frequency statistics on the keywords. As shown in Figure 10~13.



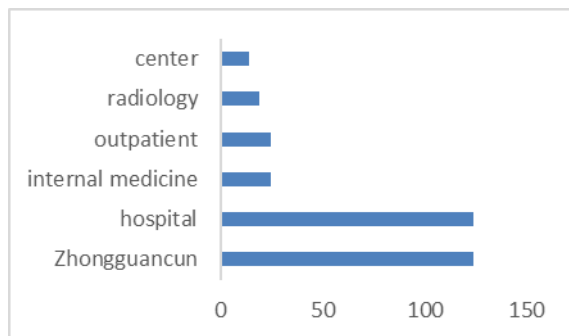
**Figure 10:** The key words frequency statistics of region R1



**Figure 11:** The key words frequency statistics of region R2



**Figure 12:** The key words frequency statistics of region R3



**Figure 13:** The key words frequency statistics of region R4

From Figures 10 to 13, it can be seen that in the R1 area given the “Home Address” label, there are keywords such as “aerospace community”, “community” and “property”. In the R2 area given the “work place” label, there are keywords such as “aerospace”, “building”, and “bank”. There are “walking street”, “Carrefour”, “Starbucks”, “hamburger” and “hospital”, “hospital”, “outpatient” and other keywords in the R3 and R4 areas that have been given “entertainment places or other locations”. Based on the above information, it is presumed that the user lives in the community of a space unit in Zhongguancun and works in the space unit. During his leisure time, he often appears in nearby Zhongguancun walking street and the hospital.

## 4 CONCLUSION

In this paper, the clustering method is used to attempt to identify personal important locations in GPS data of mobile users. In combination with reverse geocoding and natural language processing techniques, the hidden user personal characteristics are further excavated, and specific algorithm steps and technical routes are proposed. The improved density peak clustering algorithm in this paper works well on low-dimensional datasets in UCI data sets. Since the Geo\_life dataset has no a priori reference standard and there is still a reason for personal privacy, this paper cannot give specific indicators such as the recognition accuracy rate on this data set. However, the key words given by the natural language processing method have qualitatively verified the recognition results, which shows that the ADPC algorithm has certain feasibility in location data mining. For the above limitations, GPS data of volunteers can be collected in the future, and volunteers can score the recognition

results of the method as a basis for judging and recognizing the accuracy rate.

## REFERENCES

1. Nurmi P. Identifying meaningful places[D]. Helsinki, Finland: University of Helsinki, 2009:50-105.
2. Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks [C] //Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011:1082-1090.
3. Yang P, Zhu T, Wan X, et al. Identifying Significant Places Using Multi-day Call Detail Records[C]//2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2014:360-366.
4. Montoliu R, Blom J, Gatica-Perez D. Discovering places of interest in everyday life from smartphone data[J]. Multimedia tools and applications. 2013, 62(1): 179-207.
5. Hoh B, Gruteser M, Xiong H, et al. Enhancing security and privacy in traffic-monitoring systems[J].Pervasive Computing, IEEE. 2006, 5(4): 38-46.
6. Krumm J. Inference attacks on location tracks[J]. Pervasive Computing. 2007, 74(2): 127-143.
7. Zang H, Bolot J. Anonymization of location data does not work: A large-scale measurement study[C]//Proceedings of the 17th annual international conference on Mobile computing and networking. ACM,2011: 145-156.
8. Alex Rodriguez, Alessandro Latio. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492-1496.