Discovering Urban Social Functional Regions Using Taxi Trajectories

Ke Fan*, Daqiang Zhang*, Yunsheng Wangˆ , Shengjie Zhao#
*School of Software Engineering,Tongji University, Shanghai, China
Ĉomputer Science department,Kettering University, Flint, Michigan, USA
#School of Electronic and Information Engineering, Tongji University, Shanghai, China
Email: shengjiezhao@tongji.edu.cn

***Abstract*—Social function refers to the recognized human use of urban land. Detailed urban land social function identification is an integral part of urban planning. This paper focuses on discovering the social functions of urban hot areas by using taxi trajectory data. The dataset is collected in Shanghai, China, over 1.5 years. Firstly, we propose a new method to find the urban hot areas. Secondly, we classify these areas into our predefined 10 kinds of social function by depicting the characteristics of them explicitly. Thirdly, we analyze the various temporal features of ten classes in detail.**

***Keywords*-hot areas, social function, taxi trajectory**

I. Introduction

With the rapid development of a city, various functional areas are emerged to meet the diverse life requirements of urban inhabitants. The functional areas come into being by artificially designed or naturally formed due to human activities [1]. Urban land identification of social function is an important issue in urban planning [2]. The ability of identifying functional areas can also provide a quick understanding of a complex city to strangers, such as tourists. Meantime, for local citizens, it can provide the information of urban operating regularity [3]. In this paper, we focus on discovering social functional urban hot areas by using taxi trajectory. Trajectories have always been regarded as the path followed by an object moving in space and time [4]. A taxi GPS (Global Positioning System) trajectory offers us much rich information on motivations, behaviours, and dynamics of urban mobility [5].

While much work related to functional areas identification has been done before, many literatures have considered a very few kinds of functional regions (e.g. resident areas and non-resident areas). Thus, the social function of urban land can not be classified in detail. Whereas, in this paper, the social functions of urban areas are partitioned into ten labeled classes, which will be detailed later. The main contributions of this paper are three-fold: 1) We put forward a new clustering algorithm to find the hot areas in the city by taking geographical impacts into account. The clustering algorithm can be applied to the situation that the number of clusters is unknown in advance. 2) We classify urban functional areas into ten classes by designing six features that is used to depict the features of these hot areas exactly. 3) We analyze the temporal variation features of trips with certain social functions in detail.

The remaining of this paper is organized as follows. Section II proposes a new clustering algorithm for extracting hot areas in Shanghai. Section III uses the taxi trajectory data to identity the social function of hot areas. Section IV shows the results of method that were verified by the Shanghai taxi dataset. Finally, Section V concludes the paper.

II. Discovering urban hot areas

Figure 1 illustrates the framework of discovering the social functions of urban hot areas, the processing can be detailed as follows. We firstly partitioned the real map into many polygons by using Voronoi segmentation. Thus these polygons are treated as basic areas. Then, we transformed Voronoi cells into graph, which was represented by the adjacent matrix. After we turned into the question of clustering into discovering communities in graph. Next, we divided the communities obtained from graph into many area clusters. In this way, we can find the urban hot areas through estimating the area density. Six features are designed by us to depict the features of these hot areas explicitly. These features can be used to classify urban hot areas into several kinds of social function.

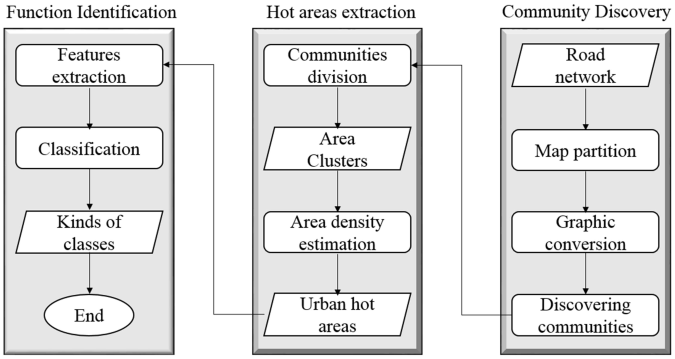


Figure 1: The framework of discovering the functional areas

There are three innovation points of our clustering algorithm: 1) We take the geographical information into consideration by dividing the whole city map into Voronoi

Table I: The list of notations

Notations	Explanation
p	an intersection position, $p = (x, y)$, x and y denote the latitude and longitude of p respectively.
N	the intersection set of p, $N = \{p : p = (x, y)\}$
$level(N)$	the degree of intersection p, ranges from 1 to 3 based on their important intersection. When $level(p_i) = 1$, the p_i is the most important intersection.
K	the key points set, which contains all key points k_i.
V_j	the Voronoi cell, $v_j = \{K_j : K_j = \{x_j, y_j\}\}$, which means the only key intersection which belongs to the Voronoi cell v_j.
$A(V_j)$	the neighborhood of Voronoi cell.

cells. 2) We add the similarity of the neighbourhoods to the measurement criteria of the same cluster on the foundation of traditional distance criteria. 3) Our clustering algorithm can be used to cluster when the number of clusters is unknown.

A. Terminology and Notations

Before further describing the method mentioned above, we introduce several definitions and notations that are used in the method. The list of notations are showed in Table I.

Definition 1: The similarity S_{ij} of the two adjacent Voronoi cells v_i and v_j can be calculated as formula 1.

$$\begin{aligned} S_{ij} &= getDist(PDR_i, PDR_j) \\ &= \frac{\sqrt{\sum_k (PDR_{ik} - PDR_{jk})^2}}{\sqrt{\sum_k {PDR_{ik}}^2}\sqrt{\sum_k {PDR_{jk}}^2}} \end{aligned} \tag{1}$$

Where PDR_i, PDR_j represent the number of pick-up points per hour in i_{th} and j_{th} region respectively. $getDist()$ denotes compute the Euler distance between PDR_i and PDR_j.

Definition 2: Let e_{ij} be the fraction of edges in the network that connect vertices in cell V_i to those in cell V_j. $adjacent(i, j) = 1$ means there are public edges between V_i and V_j.

$$e_{ij} = \begin{cases} \dfrac{1}{2m}, if\ adjacent(i,j) = 1 \\ \quad\quad 0 \quad\quad\quad\quad , others \end{cases} \tag{2}$$

Where m is the number of graph's sides, k_i is the degree of node i.

Definition 3: The parameter $a_i = \sum_j e_{ij} = \frac{d_i}{2m}$, which means the sum of every row or column.

Definition 4: The modularity Q can be computed as equation 3. It can be used to test whether a particular partition is meaningful.

$$Q = \sum_i (e_{ii} - a_i^2) \tag{3}$$

Where $\sum_i e_{ii}$ is the sum of diagonal elements. When there are no more intra-community edges than which would be expected by random chance, $Q = 0$. Furthermore, the upper limit of Q is 1.

Definition 5: ΔQ denotes the change in Q upon joining two communities. The expression of ΔQ is as follow:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j) \tag{4}$$

B. Map Partition

In order to divide the city into several cells reasonably, we first derived real road network from OpenStreetMap, then used Voronoi segmentation to divide the city.

In the OpenStreetMap, the detail information of whole urban road network is labeled by level 0, 1, 2 based on the importance degrees. For instance, the freeways, city expressways, and urban arterial roads are associated with a road level 0, 1, and 2 respectively.

The Voronoi diagram is based on closeness to key points in a specific subset of the space. Namely there is a key discrete point in each subset, and the nature of every subset can be represented by character of the point. Thus, we treated major intersections in urban transportation as key discrete points of Voronoi cells. Besides, the high-traffic regions are described by smaller cells, and the low-traffic regions are instead characterized by larger cells and a more coarse level of segmentation.

After we partitioned the whole urban map into Voronoi cells, we transformed them into graph, which is denoted by the adjacent matrix. The processing of building graph is in Algorithm 1.

Algorithm 1 Building Graph with Weights

```
Input:
    Intersection Set $\{N_i\}$
Output:
    Graphs with Weight G;
while $\{level(N) \leq 2\}$ do
    $K_j \leftarrow Pick\{N\}$;
end while
$V_j \leftarrow Voronoi\{K_j\}$;
for $K = 1; V_k \in A\{V_j\}; K++$ do
    for $(j = 1;\ ; j++)$ do
        Weights $S_{jk}$;
    end for
end for
```

C. Community Discovery

In this section, we used Fast-Newman algorithm to discovering the community of graph. The community structure widely exists in many networks. That means groups of vertices within which connections are denser but between

Algorithm 2 Discovering Community

```
Input:
    Weight Matrix $M$, Node $n_i$
Output:
    List of cluster $\{C_i\}$
 1: Initialize: $cluster_i \leftarrow n_i$, $k \leftarrow 1$;
 2: while $size(cluster) > 1$ do
 3:     for $i = 1 : n$, $j = 1 : n$ do
 4:         while $e_{i,j} \neq 0$ do
 5:             $cal\{\Delta Q\}$; //calculate the value of $\Delta Q$
 6:         end while
 7:     end for
 8:     $find\,\mathrm{argmax}(\Delta Q)$;
 9:     $OperationCount \leftarrow merge(C_i, C_j)$;
10:     $Count \leftarrow count + 1$;
11: end while
12: while $Count > 1$ do
13:     $i, j \leftarrow O_{count}$;
14:     $split(clustering)$;
15:     $cal\{Q\}$; //calculate the value of $Q$;
16:     $Count = Count - 1$;
17: end while
18: $find\,\mathrm{argmax}\,Q$;
19: List of cluster $\{C_i\}$;
```

which they are sparser. The concrete algorithm flow is presented in Algorithm 2.

The algorithm can be partitioned into two parts: merging and division.Steps 2-13 are the merging process, we initialized the graph to n communities, then merged communities which have pubic edges among them. We calculated ΔQ of the merged communities. As the result, we obtained one big community. Besides, the community with the biggest ΔQ will be merged prior. However, we cannot obtain the result expected from the integrated merging graph, so we must split it. We split the community with lower values of Q, which is used to test whether a particular division is meaning.

D. Hot Area Extraction

In this section, a simple method is used to divide these clusters into new clusters. Since the clusters obtained from the method above are geographically dispersed. In a word, we merge these adjacent cells based on their geographical location. Namely there are public edges among them. In this way, we ended up with a large number of clusters. In order to pick hot areas from these clusters, we estimate their cluster density based on their number of pick-up points and drop-off points. Thus, we extract the pick-up points and drop-off points first. Then we ranked them based on their number of pick-up points and drop-off points. The areas which have the value of PDRs per hour greater than 0.2 are defined as hot areas.

III. Social function identification

Social function of urban land can be defined as the recognized human use of urban land. In this section, we verify the social function of hot areas. We use the temporal and spatial dynamics of the pick-up/drop-off points to characterize these hot areas.

First of all, we provided a fine-grained classification of the social function of hot areas. There are ten kinds of social functions are defined in this paper, including: 1) resident district; 2) commercial regions; 3) leisure places; 4) working areas; 5) schoolyard; 6) scenic spot; 7) station; 8) hospital; 9) entertainment zone and 10) Industrial district.

A. Features Extraction

A good feature will be very helpful for the land-use classification of a region. There are six features designed to depict the characteristics of these hot areas explicitly. All the features are computed using the historical data of a certain time length.

The six features are listed as following: 1) $N_p \rightarrow$ The number of pick-up points per hour in a day. 2) $N_d \rightarrow$ Similar to the $N(p)$, which denote the number of drop-off points per hour in a day. 3) $\Delta N_p \rightarrow$ The change of the number of pick-up points per hour in a day. 4) $\Delta N_d \rightarrow$ The same as $\Delta N(p)$. 5) $\Delta N_{p,d} \rightarrow$ The difference of the number of pick-up points and drop-off points in each hour. 6) $R_{p,d} \rightarrow$ The specific value of the number of pick-up points and drop-off points in each hour.

B. Classification

In the last Section II, we ended up with many urban hot areas. However, the social function of these hot areas is still uncertain. For building a training set with social function, we adopt the manual labeling strategy. Only those areas with relatively single social function were labeled. In this way, we finally obtained some areas with a labeled social function. These areas belong to ten kinds of social function, which has mentioned above. In addition, the linear-kernel Support Vector Machine (SVM) classifiers is used to classify urban land-use.

IV. Experiments

A. The Dataset

The dataset is one of the few datasets that contains a exceeding large urban area, that is the city of Shanghai, the biggest metropolis in China. The dataset collects real motion traces from about 13854 operational taxies in the whole Shanghai, during 1.5 years.

The compositions of the GPS messages are including: Taxi ID, Location (longitude and latitude), Time-stamp, Instant Speed, Heading, GPS State, Loading State (from 0 to 3).

B. Preprocessing

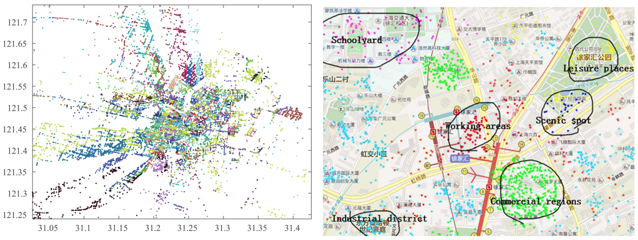
Since there are a mass of noises contained in GPS messages, e.g. the taxis are not always report their locations with same time intervals and frequency.

In this section, we utilized linear interpolation to obtain the more accurate locations. We insert location points to all the taxis to ensure a consistent time interval, every 12-s interval.

C. Results

In order to exhibit distinctly, hereby we only use the result of $400km^2$ areas in Shanghai urban center as an example.

As described above, after we merge these adjacent cells in geographical location, we got 2,059 clusters. By the way, there are 468 clusters in the $400km^2$ areas of Shanghai urban center, which is exhibited Figure 2. The points in the figure are the pick-up points, and each color denotes a communities. Besides, as showed in right figure, we take the Xuhui area in Shanghai as an example to show the result in detail. After that, we end up with 64 urban hot areas.



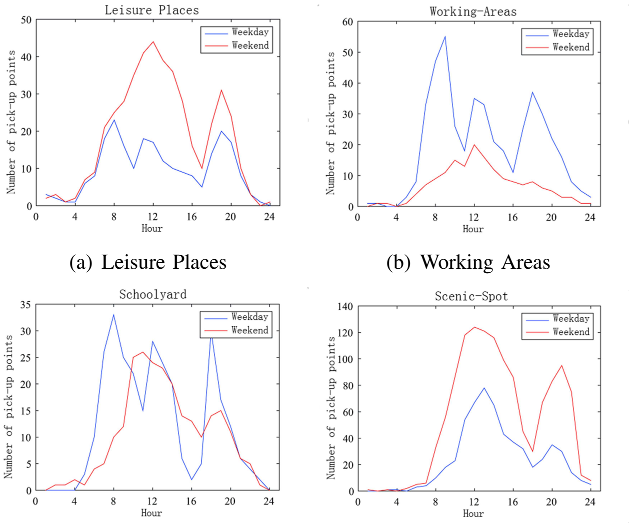
(a) The distribution of clusters (b) An example of clusters

Figure 2: The results of clustering

Then we asked ten excellent taxi drivers to manual label 429 areas, which is used as training set of SVM classifiers. Moreover, Figure 3 takes the various temporal variation features of four classes as example. The figure shows the mean trips number in each hour of a day averaged over a year for each region. In which, we explicitly separate weekdays and weekend, that is colored with red and blue respectively. Meanwhile, we classify the trips according to the class of their pick-up points. It is obvious that different functional classes differ greatly in peak value, daily fluctuation. In addition, there are some differences existing in weekday and weekend.

V. Conclusion

In this paper, we have proposed a method to discover urban hot functional areas using taxi trajectories captured in Shanghai project, which collects 13,000 taxi traces over 1.5 years. We firstly discover hot areas and then identify social function for each hot area. To be specific, we have classified the wholistic Shanghai road network into 2,059 clusters, and extracted 836 hot areas eventually. Then, we have leveraged the SVM classifier to assemble the hot areas into 10 classes.



(a) Leisure Places (b) Working Areas

(c) Schoolyard (d) Scenic Spot

Figure 3: The temporal variation of the number of trips in each class

Acknowledgment

This work is partially supported by the National Science Foundation of China (Grant No.61472283, 61471267 and 61103185), the Fok Ying-Tong Education Foundation, China (Grant No. 142006), the Fundamental Research Funds for the Central Universities (Grant No. 2013KJ034) and the National Basic Research Program of China (973 Program, No. 2014CB340404). This project is also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] J. Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

[2] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li. Land-use classification using taxi gps traces. *Intelligent Transportation Systems, IEEE Transactions on*, 14(1):113–123, 2013.

[3] D. Zhang, H. Huang, M. Chen, and X. Liao. Empirical study on taxi gps traces for vehicular ad hoc networks. In *Communications (ICC), 2012 IEEE International Conference on*, pages 581–585. IEEE, June 2012.

[4] D. Zhang, V. Athanasios and H. Xiong. Predicting location using mobile phone calls. *ACM SIGCOMM Computer Communication Review*, 42(4):295-296, 2012.

[5] Y. Zhu, Y. Wu, and B. Li. Trajectory improves data delivery in urban vehicular networks. *Parallel and Distributed Systems, IEEE Transactions on*, 25(4):1089–1100, 2014.