Urban Mobility Mining and Its Facility POI Proportion Analysis based on Mobile Phone Data

Rong XIE[1, a], Chao GONG[2, b]

[1] International School of Software, Wuhan University, Wuhan 430079, China

[2] International School of Software, Wuhan University, Wuhan 430079, China

[a]email: xierong@whu.edu.cn, [b]email: gcdofree@gmail.com

Keywords: Mobile Phone Data; Cluster; Urban Mobility Mining; Point of Interests (POI); Facility POI Proportion Analysis

Abstract. It is very important to analyze urban mobility patterns and facility distribution for urban planning and management etc. Using mobile phone data, original base station data are made statistics and analysis. We improve AHC hierarchical clustering algorithm to discover community structure of urban base station network to analyze distribution of urban mobility. We also propose an approach to regional POI proportion analysis via urban POI data. According to structure measurement indexes of city landscape ecosystem, we analyze composition and distribution of urban mobility and its influence on distribution.

Introduction

It is very important to analyze social behavior patterns as well as urban mobility distribution for some applications, like urban planning, urban construction, traffic management etc. Now mobile phones have become indispensable in our lives. Almost everyone has his/her mobile phones. With surge of amount of mobile phone users, telecom operators have accumulated large amount of mobile phone data. These large-scale data includes location information of mobile phone users. Through analyzing this location information, we can discovery some characteristics, lifestyle, and habits etc. of mobile phone users, and further spatial and temporal characteristics of group activities and distribution patterns of urban mobility. Therefore, mobile phone-based data mining is becoming an important new mean for analysis of group activity and urban mobility. Currently, some researchers propose some methods on it [1-7]. However, the existing studies do not relate to POI, as well as lack in analysis of relationship between urban mobility and regional POI proportion. Thus, we focus our research on analysis of group activity patterns, urban mobility, and regional POI proportion. The paper is organized as follows. First, it presents an approach to base station regional segmentation. And then, it analyzes major community structure inside urban base station network and urban mobility distribution through improved AHC clustering algorithm. The paper also presents proportion relationship analysis among urban facilities. Conclusion and future work are finally given in the paper.

Regional Segmentation of Base Station

We handle preprocessing on the raw data to generate some relevant base station data, such as base station location, and capacity of flow of base station. Using Thiessen polygons, urban areas are divided into several sub-areas and can be mapped on map.

a. Original data

Raw data in our research are from mobile users' calling logs in Kunming City, China. Corresponding data structures are shown in Table 1. In mobile phone base station location data, the entire area is divided into some large areas *LAC*. *LAC* is divided into several base station cells, which is subdivided into several cell sites, corresponding to large area ID (*LAC*), base station ID (*Cell_ID*) and sector ID (*Site_ID*). When mobile phone user switches between two adjacent base stations, requests of cut-in and cut-out are generated. Sum of base station cut-in and base station

cut-out represents flow between base stations. Exact latitude and longitude coordinates can be obtained by large area code (*LAC*) and cell ID (*CI).*

Table 1: Original data file structure of base station

Field	Description
time	communication time generated by base station, with units accurate to seconds
LAC	ID of large coverage area
CI	ID of base station covering a large area
NRCOC	number of requests of cut-out between adjacent base stations (e.g. base station A to station B)
NRCIC	number of requests of cut-in between adjacent base stations (e.g. base station B to station A)
NRSCOC	number of requests of successful cut-out between adjacent base stations (e.g. base station A to station B)
NRSCIC	number of requests of successful cut-in between adjacent base stations (e.g. base station B to station A)

b. Raw data preprocessing

There are seven original data types as shown in Table 2. Some data are needed to be converted, such as large area code (*LAC*), cell identification (*CI*). Exchange capacity is generated through number of requests of cut-in and cut-out. Some auxiliary data are also needed to generate, such as unique *ID* of a base station. In order to ensure complete information for further analysis, some missing or incomplete records in the original data are removed, such as number of missing cut-in requests, large area code *LAC* etc.

Table 2: Base station data structure after data processing

Field	Description	Datatype
ID_A	ID of base station A	*string*
longitude_A	longitude of base station A	*double*
latitude_A	latitude of base station A	*double*
ID_B	ID of base station B	*string*
longitude_B	longitude of base station B	*double*
latitude_B	latitude of base station B	*double*
handover	exchange capacity between base stations	*integer*

c. Base station region segmentation based on Thiessen polygons

In order to analyze group activities and regional POI, it is required to make segmentation for urban areas. Among some regional division calculation methods, Thiessen polygon method [8] considers number of sites and location of site, and accuracy is also better, so urban region segmentation is used in the paper.

To generate Thiessen polygons, firstly, determine triangle circumcenter with vertex of base station for each base station. Then, construct a Thiessen polygon by connecting to these circumcenters in clockwise. After finding corresponding Thiessen polygons of all base stations, the entire area is then covered by Thiessen polygon mesh, and thus urban region is segmented. Thiessen polygon mesh on Google map generated by this algorithm is shown in Figure 1. The implementation code of Thiessen polygons is described as follows.

```
public void addNewSite (Site site)
  Triangle triangle = locate(site); // To generate Thiessen triangle corresponding a base station
  // If the result of Thiessen triangle is empty, then report an exception.
  if (triangle == null) throw new IllegalArgumentException("No containing triangle");
  // To generate the corresponding Thiessen polygons
```

Set<Triangle> cavity = getCavity(site, triangle);
// To update the entire Thiessen polygons set, and update the current Thiessen triangle
mostRecentUse = update(site, cavity);

Fig.1. Thiessen polygon based on base station

Urban Mobility Cluster Analysis

In complex network which is composed of multiple nodes, we can divide the whole network into several different community structures through relationship among nodes. Community structure [9] in network represents a group having the greatest similarity but different from the others in network. Network community has the following characteristics. Most nodes in associations have the same or similar attributes, and relationship among nodes is also very close. Among communities, their relationship is relatively sparse. Based on clustering base stations, network is divided into a number of different community structures. Taking use of exchange capacity among base stations as a classification attribute, we can find distribution of urban group activities and urban hotspots.

a. Distribution statistics of exchange capacity among base stations

Locations of urban base stations are close. If conventional Euclidean distance calculation method is used for clustering base stations, similarity among base stations cannot be reflected. Therefore, it is considered to calculate exchange capacity between base stations for cluster index. Base stations with similar exchange capacity are clustered together. Then, combined with Thiessen polygon, analysis of clustering and distribution of urban mobility is handled.

In order to study on urban mobility, we determine the scope of urban areas in our research as the region of the Second Ring in Kunming City. The raw data is preprocessed. Filter those data whose latitude and longitude coordinates are outside the Second Ring, and append a field of exchange capacity for base station data within the Second Ring. Exchange capacity can be determined by the average of exchange capacity of all connected base stations. Class *DataPoint* is used to represent a sampling point of base station. The main attributes of this class is described as follows.

private string dataPointName; // name of sampling point
private Cluster cluster; // cluster of sampling point
private double dimension[]; // dimensions of sampling point
private double handover;// exchange capacity of sampling point
private arrayList<integer> surroundDP; // ID queue of other sampling point adjacent to the sampling point

b. Urban mobility analysis based on improved AHC algorithm

Considering characteristics of urban base station data, i.e. unknown clustering shape, large noise quantities, and moderate size of data, we improve AHC (Agglomerative Hierarchical Clustering) algorithm [10] and make clusters for base station data. The main idea is as follows. Each sampling point is regarded as an initial cluster and merge clusters to implement the final clustering. And then search for distribution of urban mobility from these clustering results.

After generating sampling point data of base station, we initialize clusters. Here, class *Cluster* represents a cluster, whose main attributes are described as follows.

private string clusterName;//name of a cluster
private double averageHandover;//average exchange capacity of a cluster

In class *Cluster*, *dataPoints* stores new generated sampling points during the clustering process.

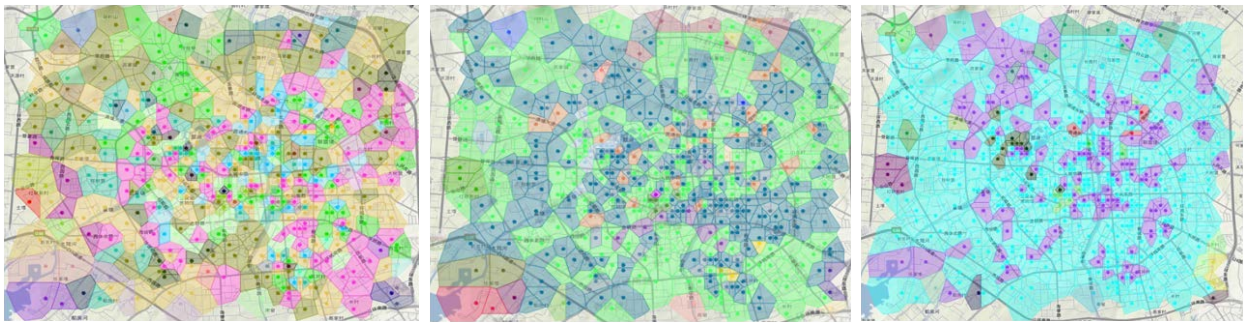
clusterName represents a unique cluster. *averageHandover* is average exchange capacity of revised data cluster during merging clustering process.

After all clusters are initialized, we can begin clustering for base stations. In the paper, cluster is handled according to exchange capacity between base stations. Thus traditional AHC clustering algorithm is required to make improvement. Exchange capacity is used for clustering instead of the use of Euclidean distance. Algorithm code is described as below for class *ClusterAnalysis*.

```
for (int i = 0; i < finalClusters.size(); i++)
   Cluster clusterA = finalClusters.get(i);
   List<DataPoint> dataPointsA = clusterA.getDataPoints();
   for (int m = 0; m < dataPointsA.size(); m++)
      DataPoint tempA = dataPointsA.get(m);
      ArrayList<Integer> surroundDP = tempA.surroundDP;
      for (int j = 0; j < surroundDP.size(); j++)
         int su = surroundDP.get(j);
         for (int k = 0; k < finalClusters.size(); k++)
            Cluster clusterB = finalClusters.get(k);
            if (k == i) continue;
            for (int l = 0; l < clusterB.getDataPoints().size(); l++)
               DataPoint tempB = clusterB.getDataPoints().get(l);
               if (tempB.getDataPointName().equals(su + ""))
                  if (Math.abs(clusterA.getAverageHandover() - clusterB.getAverageHandover())
                  / clusterA.getAverageHandover()<rate)
                     mergeIndexA = i;
                     mergeIndexB = k;
                     finalClusters = mergeCluster(finalClusters, mergeIndexA, mergeIndexB);
```

c. Cluster results

An important factor of evaluation of clustering results is the threshold of difference of average exchange capacity among clusters, which represents the only condition for merging two adjacent clusters. If difference of average exchange capacity between two clusters is less than the threshold, then these two clusters can be merged; otherwise, two clusters are retained. By observing clustering results of base station data under circumstances of different thresholds, we can find the best clustering results. Figure 2 shows Thiessen polygons-based visualization results for three situations, proportion of difference of the average exchange capacity among clusters and threshold of 20%, 40% and 50%, respectively. Patches with different colors are corresponding to different clusters.



a. Proportion of difference of the average exchange capacity among clusters and threshold is 20%.

b. Proportion of difference of the average exchange capacity among clusters and threshold is 40%.

c. Proportion of difference of the average exchange capacity among clusters and threshold is 50%.

Fig.2. Visualization of urban mobility cluster analysis

Regional POIs Analysis

After we obtain distribution of urban mobility, we wish to identify causes of phenomenon of these distributions. Here we introduce urban facilities proportion (POI), that POI is the acronym for Point of Interests, such as hotel, retail, residential, parking and other specific locations, including latitude coordinate, longitude coordinate, name, category and other geographic information. Taking

advantages of base station data and urban POI data, we can achieve POI analysis of city facilities on urban mobility distribution based on several important indexes of spatial structure measurement of urban landscape ecosystem and obtain distribution and composition of urban facilities POI, which reflecting their effects on city functional areas and city mobility distribution.

a. POI data preprocessing

Select some necessary attributes from the original POI data to generate object *POIObject*, which is described in Table 3. When generating object *POIObject*, original data is required to be filtered because of the existence of missing data and duplication data to avoid their impact on POI analysis.

Table 3: Object *POIObject*

Field	Description	Data type
ID	ID of POI object	*string*
name	name of POI object	*string*
category	category of POI object	*double*
longitude	longitude coordinates of POI object	*double*
latitude	latitude coordinates of POI object	*double*

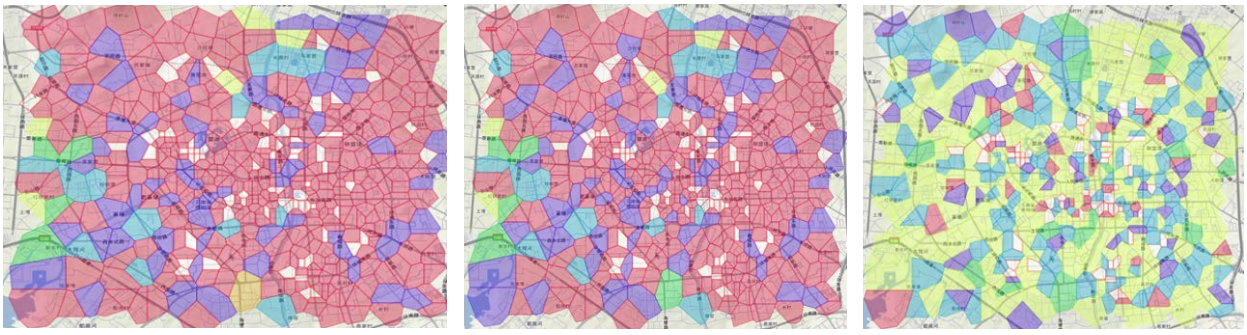
b. Spatial structure measurement analysis of urban landscape ecosystem

Urban area has a large number of different types of POI, similar to landscape ecosystem which have many different landscapes in an enclosed space. Therefore, we can introduce structure measurement indexes of landscape ecosystem for regional POI analysis. The main metrics [11] includes diversity index, evenness index and dominance index. Diversity index represents number of types of landscape and changes of percentage of each type of landscape. Evenness index describes uniformity degree of distribution of components in landscape. Dominance index indicates the maximum deviation among landscape diversities, representing degree that one or several landscape types dominate the whole landscape in landscape compositions.

c. Implementation of proportion relationship analysis of urban facilities

First, make statistics for POI with Thiessen polygon of each base station. After reading POI data, make statistics on number of categories. The implementation codes are described as follows.

```
LinkedHashMap<string, integer> poiHashMap = new LinkedHashMap<string, integer>();
for (int i = 0; i < poiList.size(); i++)
  POIObject tempObject = poiList.get(i);
  if (poiHashMap.containsKey(tempObject.getCategory()))
    // If there is the same type, then add the number.
    int value = poiHashMap.get(tempObject.getCategory());
    poiHashMap.put(tempObject.getCategory(), ++value);
  else poiHashMap.put(tempObject.getCategory(), 1); // To generate a new type.
```



a. Diversity index distribution chart b. Evenness index distribution chart c. Dominance index distribution chart

Fig.3. Visualization of proportion analysis of urban facilities

And then using these three key indexes of urban landscape ecosystem, i.e. diversity index, evenness index and dominance index, make analysis and statistics on POI data, and map them on Google map. Figure 3 shows the visualization results.

Conclusion

Through base station data processing and POI analysis, urban group activities mining and regional POI proportion analysis are handled in the paper. Main work includes three aspects as follows. 1) Preprocess raw base station data, including excluding incomplete data, filtering duplicate and redundant data, unify data format and generate available base station data. Generate base station data as Thiessen polygon network, and map them on map. 2) Regarding exchange capacity of base station as clustering index, make clusters for each base station through our improved AHC clustering algorithm. Identify community structure of base station network, and analyze urban mobility distribution. 3) Preprocess the original POI data, and generate available POI data. Using base station clustering results, Thiessen polygon mesh and relevant indexes of structures measurement of urban landscape ecosystem, we analyze urban facilities POI, including composition and structure distribution of urban facilities POI and their impact on urban mobility distribution.

Acknowledgement

This work is supported by National Nature Science Foundation of China under grant no. 41231171. The authors would like to thank Xiaoqing Zou at Kunming University of Science and Technology, Kunming, China for providing us with mobile phone data.

References

[1] Carlo Ratti, Riccardo Maria Pulselli, Sarah Williams, Dennis Frenchman. Mobile landscapes: Using location data from cell phones for urban analysis, Environment and Planning B: Planning and Design [J], 2006 33 727-748.

[2] Francesco Calabrese. Urban sensing using mobile phone network data, IBM Smarter Cities Technology Centre, 2011.

[3] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese et al. Redrawing the map of Great Britain from a network of human interactions, PLoS ONE [J], 2010 5(12).

[4] Santi Phithakkitnukoon, Giusy Di Lorenzo, Teerayut Horanont. Activity-aware map: Identifying human daily activity pattern using mobile phone data, Proceedings of HBU [C], 2010 14-25.

[5] Jonathan Reades, Francesco Davide Calabrese, Carlo Ratti. Cellular census: Explorations in urban data collection, Pervasive Computing [J], 2007 6(3) 10-18.

[6] Francesco Calabrese, Massimo Colonna, Piero Lovisolo et al. Real-time urban monitoring using cell phones: A case study in Rome, IEEE Transactions on Intelligent Transportation Systems [J], 2011 12(1) 141-151.

[7] Francisca Rojas, Francesco Calabrese, Filippo Dal Fiore et al. Real time Rome, MIT Sensible City Laboratory, 2006.

[8] Kurt E. Brassel, Douglas Reif. A procedure to generate Thiessen polygons, Geographical Analysis [J], 1979 11(3) 289-303.

[9] Paul Expert, Tim S. Evans, Vincent D. Blondel, Renaud Lambiotte. Uncovering space-independent communities in spatial networks, London: Complexity and Networks Group, Imperial College London, 2010.

[10] William H. E. Day, Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods, Journal of Classification, 1984 1(1) 7-24.

[11] Nathan Eagle, Michael Macy, Rob Claxton. Network diversity and economic development, Science [J], 2010 328(21) 1029-1031.