Research on Passenger Flow Characteristics of Rail Station Based on Mobile Signaling Data

Lilei Wang

School of Transportation and Logistics, Southwest Jiaotong University, No. 111 North Second Ring Road, Chengdu, Sichuan 610031, China
Email: wanglilei_edu@163.com

Abstract. As a kind of transport big data, mobile signaling data has been widely used in various fields of urban transport research with the characteristics of wide coverage, real-time dynamic and low acquisition costs. This paper constructs the method of extracting the travel information of the rail transit passenger flow on the basis of summarizing the shortage of the existing algorithms for recognizing the travel path of the rail transit. In the entering-station identification, signaling data generated by different signaling events with location area updating are used to identify key trajectory entering-points based on space-time sequences between adjacent key trajectory points and the location information of the rail transit cell. In the leaving-station identification, a method based on the principle of proximity was proposed which combined with the cell information near the orbital station. The results show that passenger flow characteristic information of rail transit can be obtained accurately according to the phone signaling data. The passenger flow distribution characteristics acquired can meet the nature of land use around the station.

Keywords: Urban rail transit, cell phone signaling, passenger flow characteristics

1 Introduction

It is an important basis for improving the quality of rail network service and building a high quality urban rail transit system to grasp the information of the passenger flow accurately and promptly. It can effectively grasp the passenger flow characteristics and to make the rail traffic operation more satisfied with the residents' travel demand. The use of mobile location technology to transit the travel path of residents as a new survey technology has been widely used in the travel information collection. By collecting and analyzing the mobile location data produced in the process of resident trip, the complete resident travel information can be obtained. The handset data acquisition is convenient, the cost is low, the coverage is wide and the dynamic real-time performance is strong [1]. Through the extraction and mining of the key trajectory data in the travel process of the residents' rail transit, the information of the complete resident rail traffic travel chain can be obtained. It can master the information of the passenger flow in the rail traffic and grasp the spatial and temporal distribution characteristics of the passenger flow and analyze the transit. The requirement of rail transportation operation management is provided with effective decision support.

The operation management of rail transit needs a large number of accurate passenger flow data as support. The detailed and timeliness of the passenger flow data determines the scientific and feasibility of the government decision-making department and the transit operation manager to formulate management plans. The traditional passenger traffic information acquisition mainly includes station inquiry, vehicle transiting investigation and camera investigation. There are many problems, such as small sample size, complicated investigation organization, difficult data updating and so on. In the face of increasingly complex passenger flow characteristics, especially the characteristics of transfer passenger flow between multi transit lines. The survey data are difficult to accurately reflect the characteristics of the actual traffic demand and cannot meet the needs of the expanding and complex rail transit network operation and management.

The method of traffic information extraction based on mobile location data can record individual travel chain information completely. Compared with the traditional means of traffic investigation, it has the advantages of low cost, short cycle and wide coverage, which can be more comprehensive, efficient and

STATION: the name of the transit station;
LAC: indicates the location number of the transit station;
CELLID: indicates the cell number of the orbital station.

3)Transit station location database

LINE: the name of the transit line;
STATION: the name of the transit station;
LNG: the location of the longitude of the transit station;
LAT: the dimension position of the transit station.

4) Transfer station adjacency station database

LINE: the name of the transit line;
HCZ: the name of the transfer station;
LZ: the name of the station adjacent to the transfer station.

2.1 Rail Transit Travel Path Recognition Based on Rule Algorithm

2.1.1 Passengers entering station recognition

In order to match the station signaling, from all the cell signaling databases, the individual MSID is used as the unique identification number, and all the individual signaling data Q produced in one day is extracted, and LAC_i and $CELLID_i$ are used as the retrieval objects in the individual signaling events, and each individual signaling data Qi is matched with the transit cell database GD. The LAC_m and $CELLID_m$ of the first m signaling data Q_m in Q are the same as the GD database, that is, (LAC_m, $CELLID_m$) $\in$GD (LAC, CELLID) and $EVENTID_m$=7, the signaling is the signaling data of the individual incoming stations.

Identify the station, match the LAC_m and $CELLID_m$ of the signaling Q_m with the transit cell database GD, identify the station station, record the cell information of the station, and take the M signaling data as the starting point of the transit traffic path identification.

2.1.2 Passengers exit Station Recognition

According to the station signaling, the signal data of the individual is identified on the basis of the station recognition and the signaling data Q_m of the incoming station is used as the starting point to identify the signaling data after the m signaling. When the first signaling Q_n's LAC_n and $CELLID_n$ do not belong to the underground cell system database, that is (LAC_n, $CELLID_n$) $\notin$GD (LAC, CELLID), and when $EVENTID_n$=7, the signaling Q_n is the individual outgoing signaling recording data.

2.2 Service Range Identification of Rail Station Based on Spatial Clustering Algorithm

DBSCAN is a more representative density clustering algorithm *[7-8]*. By calculating the density of points in a certain range, it divides all points into different types, and allows the fusion of point sets satisfying certain requirements. Finally, a random cluster is developed in the noisy data space. Each cluster is defined as a density phase. The maximum set of points.

Definition 1: Eps-neighborhood of a point. The area of a given object whose spatial distance is Eps.

Definition 2: Core point. If the sample points in E neighborhood of p are larger than or equal to a threshold MinPts, then p is called a core point, as shown in FIGURE 1.

Definition 3: Directly density-reachable. For the sample set D, if the sample point q is in the E neighborhood of point p, and p is a core point, then the object q is directly density-reachable from the object p.

Definition 4: Density-reachable. For a sample set D, given a set of sample points p_1, p_2,..., p_i, if p= p_1 and q= p_i, if the object p_i is directly density-reachable from p_{i-1}, then the object q is density-reachable form object p.

Definition 5: Density-connected. There is a point o in the sample set D. If the object o to object p and object Q are all density-reachable, then point p and point q are density-connected, as shown in FIGURE 1.

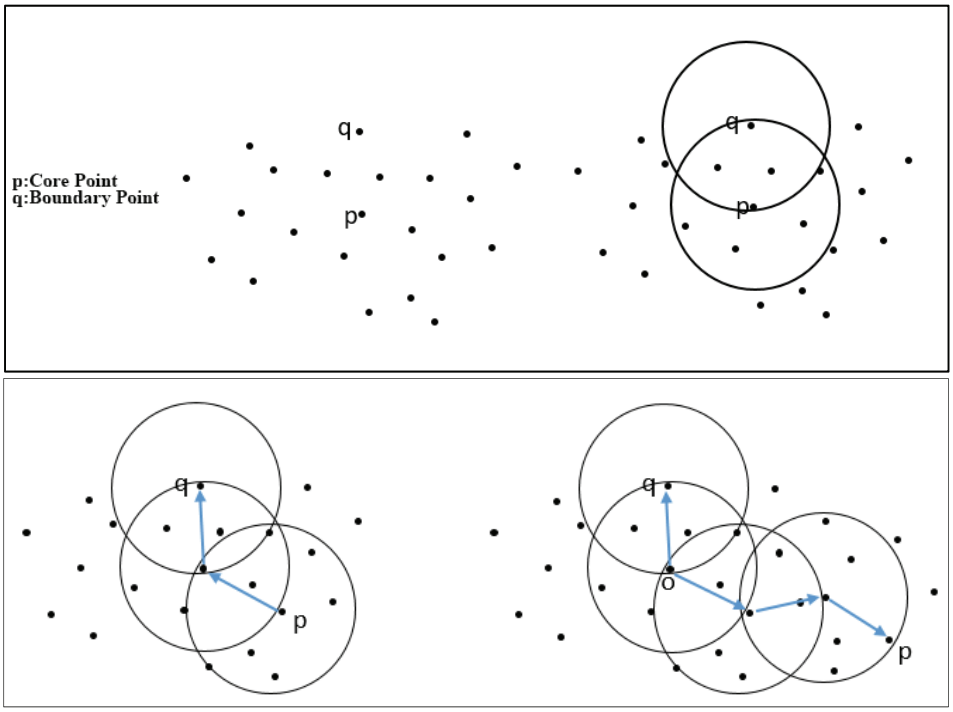

Figure 1. DBSCAN parameter and clustering mechanism.

3 Data

3.1 Data Collection

This experiment is to explore the characteristics of the signaling data produced in the course of the rail transit trip from the experimental data, excavate the regularity of the signaling data sequence, and construct the method of the transit traffic information extraction model.

Therefore, based on the requirements of modeling and analysis, this paper selects the passenger travel data of Chongqing rail transit station and analyzes the mobile signaling data around the station.

3.2 Data Pre-processing

The mobile communication system is unstable. There are some problems in the original data, such as data anomaly, data loss and the requirement of data accuracy that cannot be analyzed. It is necessary to preprocess the original data before the algorithm is processed, to "clean" the data, and to reduce the missing omission produced when the data is measured as much as possible. In order to sort out the signaling data of volunteers by time, the following data cleaning work is needed to remove invalid data. The invalid data is the data that does not record the travel behavior in the signaling data, including data such as repeated data, missing data, communication failure and so on.

(1) Eliminate duplicate data

A large number of repeated data will affect the accuracy of recognition and will also expand the amount of calculation and have a bad effect on the follow-up statistical analysis. Therefore, it is necessary to delete the data that is completely repeated.

(2) Eliminate missing data

Due to the instability of communication signals, partial field missing in partial signaling event data will interfere with subsequent calculations, such as data misplacement, and computer program reading errors. Therefore, it is necessary to delete the data of the signaling event with a field record. At the same time, the data that is not 0 of the FLAG value is also deleted because it indicates that the IMSI is not available and the user's identification number cannot be determined.

(3) Eliminating communication failure data

In the communication process, when the communication event fails, the cell also records the failed signaling data. This part of the data may not record the user's location information. Therefore, the failure data of the communication will be deleted by the event type decision of the EVENTID.

(4) Eliminate small sample data

Some users generate less signaling than the minimum requirement for analyzing travel behavior, so we delete those data with less than three CID changes in a day's LAC.

4 Verification and Result Analysis

4.1 Method Verification

A total of 295 signaling basic data of 3 groups of different transit travel conditions are selected, of which there are 1 groups each of which are no transfer trip, transfer (different station transfer) combination, and transfer (the same change) combination. Four experimental stations were selected and analyzed based on DBSCAN clustering algorithm.

4.1.1 Service range identification of rail station

Through the model analysis, 95% cell points can be clustered when the parameters are Eps=200m. When Minpts=5, the number of cells in the same land range is distributed, and clustering is the better effect. The model calibration effect is shown in Figure 2.

After the selection of the parameters, the clustering analysis can be started. The clustering process, as described by the DBSCAN algorithm, finally has 36643 points to be clustered. After the clustering is completed, 20 colors will be extracted randomly from the RGB color. Each class is represented by a color in a cycle, and the scatter plot is made, and the clustering results are obtained as shown in Figure 3.

Based on the DBSCAN clustering method, MATLAB programming is used to select parameters Eps=200m, Minpts=200m, to cluster the cells outside the station passenger stations. The clustering effect of the cell conforms to the property of the surrounding land and the setting rule of the cell. The data of each station is grouped according to the completed station clustering, and the number of signaling is added to get the destination distribution map of the station passenger flow. Figure 4 shows the visualization of the station service scope after visualization.

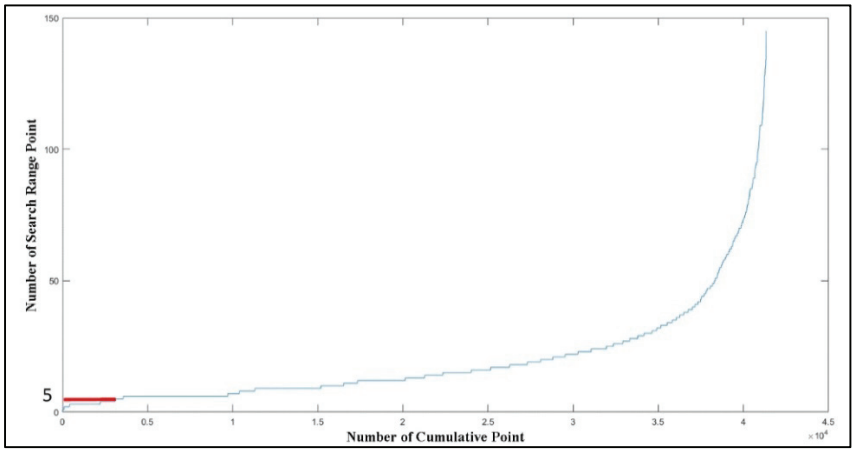

Figure 2. The results of number of search range cells clustering.

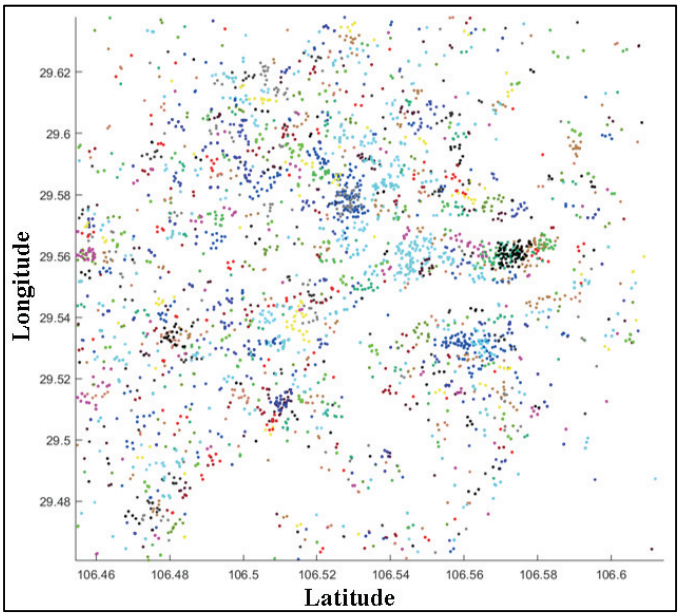

Figure 3. The clustering results of cell based on DBSCAN.

Figure 4. Thermal diagram of service range.

4.2 Result Analysis

4.2.1 Service range identification of rail station

Taking Chongqing Xiao Shi Zi station as an example, the service range of the incoming passenger flow is as shown in Figure 5. The source of the passenger flow of the station is mainly shown in the south-west direction and the low direction in the northeast. The southwest direction of the station is the main attraction area of the early rush station, the north and the south of the station have few passenger flows, and the east of the station attracts the passenger flow much less. According to the analysis of the nature of the land use around the small station, it is found that the two directions of the West and the south are mainly residential, residential and commercial land, the East and the north of the station are mainly commercial land. The greater the residential land, the higher the intensity of passenger flow, therefore, the

source of passenger flow of the station shows that the intensity of the East and the north passenger flow is obviously lower than that of the West and the south. The service range of the station early peak is basically consistent with the land use nature around the station.

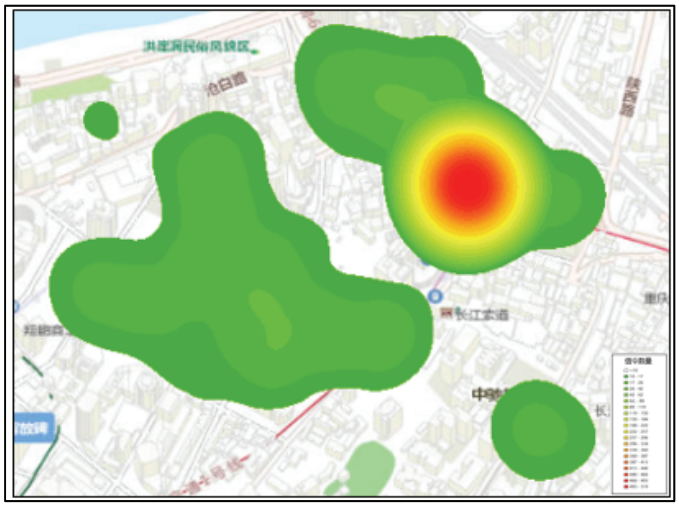

Figure 5. Thermal diagram of service range.

As shown in Figure 6, there are a number of passenger flow attractions outside the station. The east of the station is the main attraction of passenger flow, attracting more passenger flow. There are some passenger flow attractions in the West and the north, and the passenger flow is less. The station is located in the commercial center of Chongqing, the east of the station is mainly the commercial land area, the main destination area of the arrival passenger flow, and the west and the south of the station are the mixed land for commercial and residential. The overall intensity of the overall passenger flow is slightly lower than that of the eastern commercial center. With the increase of the distance from the station, the intensity of the passenger flow decreases gradually, and the distribution characteristics of the passenger flow basically conform to the nature of the surrounding land use.

Figure 6. Distribution map of outbound passenger flow.

5 Conclusions and Future Work

The number of mobile phone signaling data is large, covers a wide range, and has strong real-time dynamic characteristics. It is feasible to extract the travel information of the rail traffic, combining the layout of the cell of the rail transit system. The transit traffic service range algorithm based on DBSCAN clustering algorithm can extract the source of the incoming station, the outbound passenger flow and the feature of the direction distribution and identify the service range of the transit station.

This paper selects the characteristics of the passenger flow around the station as the research object and analyzes the passenger flow service scope of the railway station. The evaluation of the characteristics of the source of passenger flow is determined by the land development and utilization around the station. In the follow-up study, we need to select more indicators to quantify, and we can further study the characteristics of passenger flow with the more detailed questionnaire survey or other data.

Acknowledgements. This work was supported by National Natural Science Foundation of China under Grant no. 51178403, the Fundamental Research Funds for the Central Universities (no. SWJTU11CX080 and no. 2682014CX130).

Reference

1. Fei Yang, Zhenxing Yao and Peter J. Jin. Multi-mode trip information recognition based on wavelet transform modulus maximum algorithm by using gps and acceleration data. *Transportation Research Record*, 2015.
2. Systems R E, Farradyne P B. Final evaluation report for the Capital-ITS operational test and demonstration program, 2007.
3. White J, Ivan W. Extracting origin destination information from mobile phone data. Road Transport Information and Control，Eleventh International Conference on (Conf. Publ. No. 486). IET, 2002.
4. Carlo Ratti, Riccardo M. Pulselli， Sarah Williams， Dennis Frenchman. Mobile landscapes using location data from cell-phones for urban analysis. Senseable City Laboratory Massachusetts Institute of Technology, 2007.
5. Carlo Ratti, Pinelli Fabio, Hou Anyang. Space and time-dependant bus accessibility a case study in Rome. The 12th International IEEE Conference on Intelligent Transportation Systems. 2009, 10: 346-351.
6. Calabrese F, Colonna M, and Lovisolo P. Real-time urban monitoring using cell phones a case study in rome. *Intelligent Transportation Systems*, Vol. 12, No. 1, 2011, pp. 141-151.
7. Bi F M, Wang W K, Chen L. DBSCAN: Density-based spatial clustering of applications with noise. *Journal of Nanjing University*, Vol. 48, No. 4, 2012, pp. 491-498.
8. Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, pp. 226-231.

