ATLANTIS PRESS

Advances in Computer Science Research, volume 87

3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

Algorithm and Application for Labeling Workplace and Residence based on Traffic Big Data

Jie Wang [a], Yunyao Zhou [b, *]

Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, School of Information and Engineering, Wuhan University of Technology, Wuhan, China

[a] 897408159@qq.com, [b, *] zhouwhut@yahoo.com.cn

Abstract. Congestion in urban makes too many commuters choose public transport for traveling. A large number of public transportation travel data can accurately calculate workplace and residence of the regular passengers. The label of workplace and residence helps to analyze urban migration and the distribution of workplace and residence.

Keywords: The label of workplace and residence; Regular travel; Traffic big data.

1. Introduction

Nowadays, the urban population is rising sharply, and travel problems are getting worse. With the advantages of less consumption of resources, less space, and low travel cost, urban public transport has increasingly become the main solution to solve the problem of urban traffic congestion. At present, the priority development of urban public transport in China has risen to the national strategic level. Shenzhen has witnessed rapid development of public transportation in recent years. The average daily passenger volume of Shenzhen public transportation has exceeded 10 million, ranking among the top cities in China. Among them, the share rate of motorized travel in bus stations is 56%, and the public transport system is mature and convenient, which has become the primary option for people to travel. The average daily passenger volume of rail transit has exceeded 4.48 million, and the improvement of equipment quality enables a large number of high-quality user travel data to be used.

Public transport data actually records people's travel trajectories. In the case of high data quality and availability, it is a feasible scheme to use the data to calculate the label of the passenger's work place and residence, and no other algorithm based on public transport data is found at present. The public transport data used in this paper includes the data of bus and subway, which are also the mainstream travel tools for people. The label of workplace and residence refers to the bus station or subway station where a specific user is closest to his/her place of work and home address. Usually, users who choose to commute by public transportation will live near the boarding station, which is similar to their living place.

The calculation of this label is of practical significance and can be used to study the distribution of workplace and residence, the migration and the balance of a city’s workplace and residence.

2. The Relevant Data

In this paper, bus transaction data, subway transaction data and site GPS data are used. The daily data volume is about 1.1g, and the calculation of occupation and residence label is based on a month's data volume, which is more than 30G. Use Spark SQL for processing.

2.1 Metro Data of SZT

Metro data is the data when someone enters or leaves the station. Relevant fields include card number, card time, station name and card swiping type. A normal travel record includes inbound and outbound card swiping, and no other inbound data is available in the interim.

Original data need to be cleaned and cleaning rules as follows, put all the data according to user groups, each user's data according to time sequence, and the two data of two adjacent two teams, screening station - the outbound records as a OD (Origin - Destination) record, as a standard data

cleaning, filter out not matching to the OD of dirty data. Daily data volume is about 500M, with 6 million pieces of data per day.

2.2 Bus Data of SZT

The bus data is the card swiping record of each passenger, and the relevant fields have the card number and card swiping time. The original data is only the card swiping record when getting on the bus. The station information of getting on the bus should be combined with the GPS data uploaded by the bus. and the number of trips should be divided. The GPS data were divided by the trip partition algorithm to match the original bus data by license plate and time. Daily data volume is about 600M, with 5 million pieces of data per day.

2.3 GPS Information of Bus and Subway Station

Crawl the GPS information of bus station and subway station in Shenzhen to reserve. The distance between the bus stations is about 500 meters, with more stations, denser distribution, and closer to the work place and residence. The distance between subway stations is about 2km, and there are 168 stations in Shenzhen.

3. The Algorithm of Labelling

This algorithm needs to integrate the data of bus and subway and is divided into three steps. First, select the possible sites where each user lives and works. Second, the card swiping frequency of each user at each station is counted, and the bus station or subway station within 500 meters is counted as the same station. Third, by the degree of support and site weight to select the most likely user tags.

3.1 Site Selection

This algorithm is mainly aimed at those who use public transportation to commute. The accurate label has higher requirements on the accuracy of the initial selection of workplace and residence station, and the reasonable selection has great influence on the result. I'm going to define two terms here:

HS (Home-Station): the possibility of a passenger's address

WS(Work-Station): the possibility of a passenger's workplace

Here, selects the first swipe of the user's card in a day as the HS. Most users start their first trip from their home address, while a few night workers may start their first trip from their work place. Here, the travel habits of most passengers are selected as the standard.

Choose the first trip three hours after the first trip as a possible place of work. The reasons are as follows:

According to the travel time distribution in Shenzhen, a normal public transport trip can be up to 2.5h, and 3h is selected as the time interval to ensure that the travel record selected is not the same one.

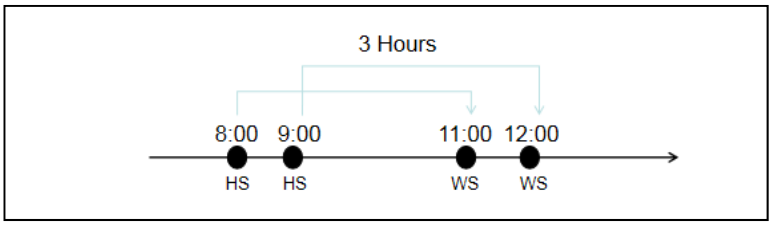

Figure 1. Distribution of the HS and WS

As shown in the above picture, after commuting passengers travel in the morning peak, another three hours is the lunch break. The intention of the passenger to travel again is as follows:

(1) Going out for lunch

(2) Go out for a meeting

(3) Go home and get something
(4) Go home for lunch break
(5) Go home after work

The starting point in the above cases is the place of work. For the above reasons, the departure site after three hours of the first trip is selected as a workplace.

3.2 Statistics of Work and Home Address Sites

According to the above rules, the travel records of passengers for 30 days will be counted. Passengers can choose to travel by bus or subway, so the result which has been chosen may be a bus or subway station. Consider that the passenger may not have a unique departure station of choice and count two stations less than 500 meters apart as the same station. Then the frequency of each site is counted. Using longitude and latitude (MLonA, MLatA) and (MLonB, MLatB), the following formula for calculating the distance between two points can be obtained according to the triangular derivation:

$$C = \sin(MLatA) * cos(MLonA - MLonB) + cos(MLatA) * cos(MLatB) \quad (1)$$

$$\text{Distance} = R * arccos(C) * \pi/180 \quad (2)$$

3.3 Determine the Label

The label is determined by taking into account two factors, one is support, the second is site weight.

Since a user's HS and WS are only once a day, the support degree needs to calculate the minimum number of days of attendance for a month for commuting users. The number of trips per month is calculated, as shown in the figure below.

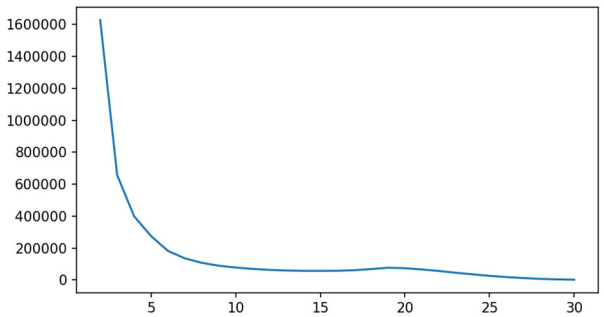

Figure 2. Passenger monthly travel frequency

It can be seen from the above figure that the largest number of users in Shenzhen are those who travel once or twice, including most users of one-way tickets and those who occasionally travel by card. In addition to the one and two peaks, it can be seen that the users who travel 17 times in the last 30 days also have a small peak, which is the commuter users who use public transportation every month. This part of users travels regularly, and the starting point and end point of travel do not change much, so the user's label can be accurately calculated.

The support degree can filter out commuter users, and the minimum weight of the station is needed to determine the label. Here, 0.8 is taken as the threshold, that is, the proportion of passenger's attendance in the station is greater than or equal to 0.8, that is, the station is determined to be the passenger's final label.

4. Label Application

The label algorithm based on the travel data of Shenzhen can label 733,653 people with the home address Tag and 438,479 people with the workplace Tag. Two applications are implemented based on the above data:

(1) Distribution of working place and living place in Shenzhen.

(2) The impact of the opening of line 11 in Shenzhen on the location and address distribution near line 1.

4.1 Distribution of Working Place and Living Place in Shenzhen .

Table 1. Address distribution table

Station	Num	Station	Num
PingZhou	36315	DaFen	12639
WuHe	21051	MinLe	11854
BuJi	19293	DaXin	11650
QingHu	19167	CaoFu	11377
BaiShiLong	18761	ChangLong	11249
DanZhuTou	17486	DaJuYuan	11167
MinZhi	16777	ShiJieZhiChuang	11053
LongSheng	16324	XiaShuiJin	10839
LongHua	16232	MuMianWan	10374
BaiShiZhou	15707	FuMin	10223
GuShu	15608	ShenZhenBeiZhan	10082
XiXiang	13872	GangXia	10057

Figure 3. Thermodynamic chart of address distribution

Table 2. Workplace distribution table

Station	Num	Station	Num
ShenDa	25226	ShangMeiLing	7976
GaoXinYuan	24410	KeYuan	7772
CheGongMiao	22685	TaoYuan	7753
DaJuYuan	20213	DaXueCheng	7502
HuiZhanZhongXin	14948	FuTianKouAn	7133
HuaQiangLu	12634	HuaXin	6823
LaoJie	9141	WuHe	6435
FuTian	8984	GangXia	6358
HuaQiangBei	8747	XingDong	6069
GuoMao	8586	ShiMinZhongXin	5340
GouWuGongYuan	8561	HuangMeiLing	5314
KeXueGuan	8256	YanNan	4778

Figure 4. Thermodynamic chart of workplace distribution

Since the longitude and latitude of each site is fixed, the thermal graph will show a point distribution. In general, the number of users whose residential addresses are in line 1, 3, 4 and 5 are mostly distributed in blocks, including BaoAnZhongXin, ShenZhenBei, LongHua, BuJi and FuTian. The work place is relatively concentrated in the two large areas around the ShiMinZhongXin and HuaQiangBei.

4.2 The Impact of the Opening of Line 11 in Shenzhen on the Location and Address Distribution Near Line 1

Calculate the label before and after the opening of line 11, track the user ID, and filter out the person who transferred from line 11 to line 1.

Table 3. Address Migration Table

Initial	migration	population
XiXiang	BiHaiWan	1015
TaoYuan	NanShan	751
JiChangDong	FuYong	530
JiChangDong	QiaoTou	416
BaoTi	BaoAn	346
DaXin	NanShan	186
JiChangDong	ShaJing	184
JiChangDong	TangWei	177
PingZhou	BiHaiWan	126
JiChangDong	MaAnShan	116

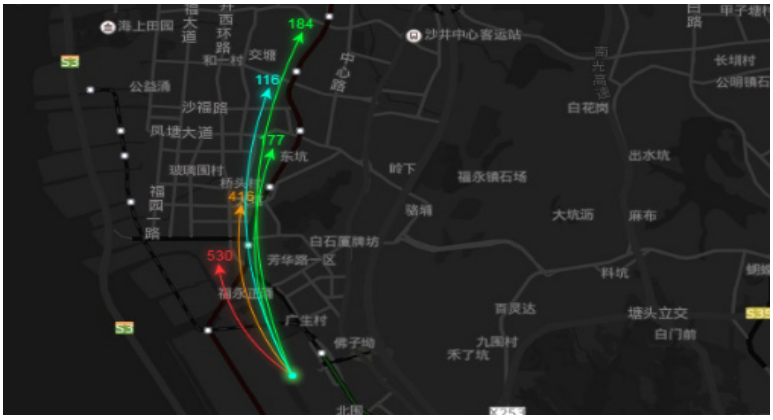

Figure 5. Address migration

Because line 11 is parallel to line 1, after line 11 opens, the situation that the user of line 1 moves to line 11 inevitably appears. The most obvious moveout situation is JiChangDong Station on Line 1. Because FuYong Station and QiaoTou Station are close to the JiChangDong Station, passengers

chose line 11 nearby. And in the top 10 statistics, a total of 1,423 users moved out. In addition, the most obvious move is the BiHaiWan Station on Line 11, with a total of 1,141 users moving in. Data show that the opening of line 11 has indeed reduced the travel pressure for line 1. In addition to the passenger flow from line 1 to line 11, the opening of the new line has also induced new passenger flow, enabling more people to enjoy convenient public transport.

5. Verification of the Algorithm

5.1 The Questionnaire

To verify the accuracy of the algorithm, we collected questionnaires near several subway stations. More than 160 questionnaires were collected. The questionnaire included the card number, the nearest subway station to the address, and the nearest subway station to the place of work. We calculate the user's workplace and residence label through the card number, and then check with the questionnaire. The data showed that 84 percent of the address labels matched, and 71 percent of the workplace labels did.

5.2 Field Research

According to the data in the fourth section, the housing clusters are PingZhou, WuHe, BuJi, etc., while the working clusters are ShenDa, GaoXinYuan, CheGongMiao, etc. We went to the vicinity of these sites for research, and through the network to look up nearby building information. The results of the algorithm are consistent with reality. The subway in the morning rush hour in the residential area is extremely crowded, while the work area is extremely crowded in the evening rush hour.

The above data show that the occupation label algorithm is effective and feasible, with research significance.

6. Conclusion

This paper introduces the algorithm of work place and residence based on Traffic Big Data, and shows the application of the algorithm. The algorithm has the advantages of relying on less data, not involving user privacy, and strong practicability. It mainly uses the data of bus and metro, and gets the label of workplace and residence accurately, which can be applied to the study of the distribution of workplace and residence, the transfer of workplace and residence, the balance of workplace and residence, etc.

References

[1]. Zhao X, Rong J. Study of the Effects of Alcohol on Drivers and Driving Performance on Straight Road[J]. Mathematical Problems in Engineering,2014, (2014-2-23), 2014, 2014(1):1-9.

[2]. Evangelia B, Moore J L, Gilbertson A D, et al. Validity of the Clock Drawing Test in predicting reports of driving problems in the elderly[J]. BMC Geriatrics, 2004, 4(1):1-7.

[3]. Evans A W. The Economics of Residential Location 1973[J]. Economica, 1973, 42(167):340.

[4]. Waddell P. Exogenous Workplace Choice in Residential Location Models: Is the Assumption Valid? [J]. Geographical Analysis, 1993, 25(1):65-82.

[5]. Benakiva M, Bowman J L. Integration of an Activity-based Model System and a Residential Location Model[J]. Urban Studies, 1997, 35(35):1131-1153.

[6]. Sermons M W, Koppelman F S. Representing the differences between female and male commute behavior in residential location choice models[J]. Journal of Transport Geography, 2001, 9(2):0-110.

[7]. Irtema H I M, Ismail A, Borhan M N, et al. Case study of the behavioural intentions of public transportation passengers in Kuala Lumpur[J]. Case Studies on Transport Policy, 2018.

[8]. Zhou J, Long Y. Bus Commuters' Jobs-Housing Balance in Beijing: An Exploration Using Large-Scale Synthesized Smart Card Data[C]// Transportation Research Board 92nd Annual Meeting. 2013.

[9]. Seifert H, Nissen V. Crowd Workplace—A Case Study on the Digital Transformation Within IT- and Management-Consulting[J]. 2018.