

Model Research on Passenger on and off Site Based on Bus IC Card and GPS Data

Shuo Wei¹, Wenyong Li^{1,2}, Ziwei Yu¹, Tao Wang¹, Anlei Yang¹

¹School of Architecture and Transportation Engineering, Guilin University of Electronic Technology, Guilin Guangxi

²Guangxi College Laboratory of Intelligent Transportation System, Guilin Guangxi
Email: 364756933@qq.com

Received: May 1st, 2018; accepted: May 15th, 2018; published: May 22nd, 2018

Abstract

To enhance the operation efficiency and management level of city bus system and improve the city traffic conditions has been the development goal of traffic management departments and bus enterprises, and also the fundamental requirement to create the Transit Metropolis. Now the bus IC card system and mobile GPS technology have become matured. They have been applied on most cities of through country. This text utilized the time correlation to match the data, and construct the fusion model of bus IC card data and GPS data, to realize the results that we can master the dynamic passengers' data of getting on or off and real-time analysis. At last, to use the bus trip data of Shenzhen, we collected to make error analysis.

Keywords

IC Card and GPS Data, Bus Station, Error Analysis

基于公交IC卡和GPS数据的乘客上下站点模型研究

魏 硕¹, 李文勇^{1,2}, 余子威¹, 王 涛¹, 杨岸磊¹

¹桂林电子科技大学建筑与交通工程学院, 广西 桂林

²广西智能交通系统高校重点实验室, 广西 桂林

Email: 364756933@qq.com

收稿日期: 2018年5月1日; 录用日期: 2018年5月15日; 发布日期: 2018年5月22日

摘要

提升城市公交系统的运行效率和管理水平,改善城市交通状况,一直是交通管理部门和公交企业的发展目标,同时也是创建“公交都市”的根本要求。如今公交IC卡系统和车载GPS技术已经成熟,在我国多数城市得到应用,本文通过时间关联匹配,构建公交IC卡刷卡数据和GPS数据的融合模型,实现了对乘客上下车情况动态掌握和实时分析,最后应用所收集到的深圳市的公交出行数据进行误差分析。

关键词

IC卡和GPS数据,上下车站点,误差检验

Copyright © 2018 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

面对城市里交通供给与交通需求之间巨大的不平衡,业内专家学者、交通管理部门和公交运营企业已经认识到,单纯加强基础设施投资建设是不够的,甚至会加剧城市的拥堵,着眼于城市公共交通事业的发展,不断改善公交出行环境,优化公交线网结构,才是解决城市交通问题的破题所在。

随着物联网技术的不断发展,公交一卡通在国内大量城市得到了广泛应用,持卡出行的居民占比持续增加,截止2015年底,深圳通卡累计发行量突破3000万张,已实现地面公交、地铁线路100%覆盖,在部分出租车上也可以刷卡。一卡通刷卡数据作为公交乘客出行行为的记录,通过与城市智能公交系统其他动静态信息的结合,经过挖掘分析后,可为公交线网规划、企业运营管理和市民出行服务提供决策依据。

王炜等提出的“逐条布线,优化成网”的方法,以直达客流量最大为目标,重点考虑单条公交线路的优化,具有很强的实践指导作用。

华南理工大学徐建闽[1]利用GPS与公交IC卡数据生成了乘客OD,其中推算乘客上车站点的原理即将GPS位置信息与IC卡数据的交易记录信息进行匹配,即将GPS中包括的位置、时间等信息与IC卡数据中的交易时间等信息联系起来。

然而将公交IC卡数据与GPS数据融合分析后,得到客流特征和线网的运行信息,此方面的研究还有不成熟的地方。本文所进行的基于公交IC卡和GPS数据挖掘分析的乘客上下车站点模型研究,模型应用可以作为公交管理决策的依据,具有一定实际意义和应用价值。

2. 上车站点识别

2.1. 乘客上车站点匹配

由于公交IC卡刷卡数据只有上车刷卡时间、所乘车辆等信息,而缺少上车时的刷卡站点位置信息,也就无法得到完整的乘客出行行为特征,因此想要获取乘客下车信息,分析乘客整个出行链行为特征,须利用GPS数据和IC卡数据融合,根据时间匹配先识别乘客上车站点的近似位置信息,再利用密度聚类算法对上车站点聚类,以此准确判断的公交站点位置。

在进行站点匹配前，首先要对站点进行层次划分。RouteStop 代表公交车在某个公交站台周围的行驶路径，Stop 代表公交停车区域，将处在一定范围内，具有连续性的 Stop 和 Routestop 称作一个 Stoparea，如图 1 所示。

按照此思路划分，深圳市共有 Stoparea 0.5 万个，全市站点 Stoparea 如图 2。

2.2. IC 卡与 GPS 时间数据修正

由于同一站点乘客往往不是单人上车，刷卡记录会有多条，但是站点 GPS 数据只有一个，公交 IC 卡所记录数据时间和 GPS 系统所记录时间往往不能够完全对应。收集深圳市 IC 卡刷卡记录 3 万条，取第 24,100 至 24,600 共计 500 条，剔除噪声数据后以 IC 卡刷卡时刻为基准时刻，统计公交车辆的 GPS 时间、刷卡时间、实际到站时间和离站时间，可以明显看出存在时间差异性，时间差异情况如图 3 所示。

这种时间差异的情况，经常会带来时间匹配时数据的缺失，为了降低影响，我们在进行时间匹配前需要先对有关数据进行修正。

假设一个修正时间差为 Δt ：

$$\Delta t = T_{\text{GPS}} - T_{\text{IC}} \quad (1)$$

式中 T_{IC} 为 IC 卡记录时间， T_{GPS} 为 GPS 记录时间。对 Δt 进行重复计算以求得最后结果，当站点准确率到达理想值后，即认为此时对应的时间差为此线路的系统最优修正时间差。按照 GPS 时间，通过最优修正时间差对 IC 卡的记录时间进行修正。

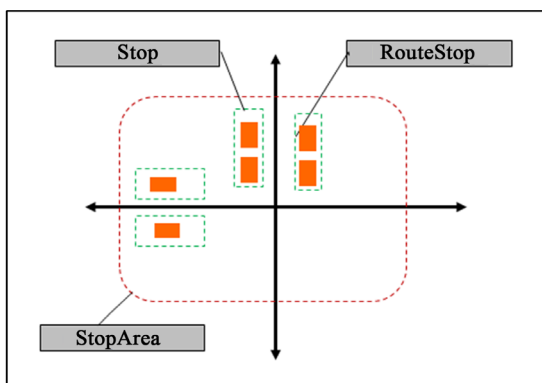


Figure 1. Description of StopArea

图 1. StopArea 说明

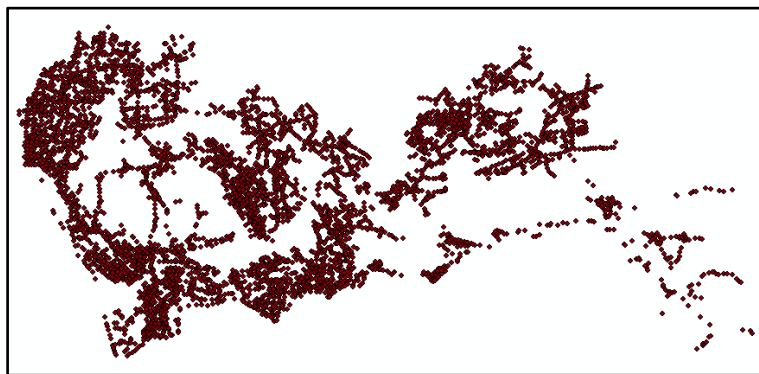


Figure 2. Summary of Shenzhen's StopArea

图 2. 深圳市 StopArea 汇总

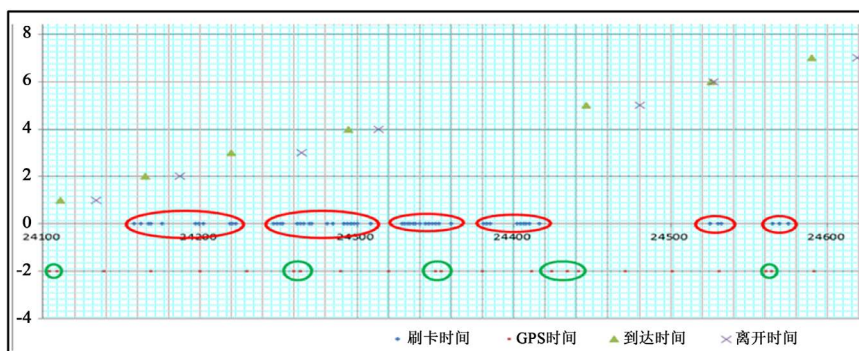


Figure 3. Difference between IC card and GPS on time data

图 3. IC 卡与 GPS 数据时间差异说明

2.3. 上车站点的聚类分析

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 聚类算法, 是一种基于密度的聚类算法, 依据高密度连通性区域, 可以聚集起密度足够高的区域, 并且在噪声中发现数据中任何形状的簇。相连点的最大密度集合区域划分为簇, 簇的聚类在有噪声的空间数据中可以以任意形状出现。由于本文先对 IC 卡数据和 GPS 数据使用时间匹配, 因此难免会形成数据噪声, 会对站点识别结果的精确度产生不利影响, 所以本文采用 DBCSCAN 密度聚类算法可以降低数据噪声对站点识别的影响, 提高聚类的精确度[2]。

Step 1: 获取 IC 卡刷卡记录的起止时间, 获取 GPS 到站记录的起止时间, 取两个时间的交集并外延 30 min, 得到起止时间 t_0, t_1 ;

Step 2: 获取 t_0, t_1 时间区间内的所有到站时间记录, 划分起止时间;

Step 3: 读取一个车次的所有刷卡记录 A, 记每一条记录为一类, 则共有 n 类, 刷卡记录均是按照事件顺序排列的, 因此取连续的两类进行聚类, 计算它们的时间间隔, 重复此步骤直到找到最小时间间隔时停止聚类, 得到聚类结果 t_a , 各站点聚类结果集合记为 C;

Step 4: 读取一个车次的所有到站数据 B, 到站时间记为 t_b , 若 $|t_a - t_b| < 10 \text{ min}$, 则将 C 全部平移 $|t_a - t_b|$ 生成时间记录 D, 用记录 D 匹配到站时间记录 B, 计算匹配相似度 v_n ;

Step 5: 若 $|t_a - t_b| \geq 10 \text{ min}$ 成立, 则求 v_n 中的最小值, 确定最佳平移时间, 继而确定最佳匹配下车站点, 记录并保存。

3. 乘客下车站点推算

分析乘客的出行情况以及获取车内乘客的数量, 对于掌握整个公交系统的运行态势具有重要意义。由于国内的公交普遍采用一票制, 公交 IC 卡只记录了乘客上车刷卡时的信息, 缺失了乘客的下车信息, 无法直接根据站点数据、IC 卡数据和 GPS 数据精确地得到乘客实际的下车站点。本文根据出行链的思想, 设计了判别乘客下车站点的算法, 根据乘客在各个站点的下车可能性推断其最可能下车站点。

3.1. 站点权重

乘客高频站点吸引权(频次): 收集某乘客在一段时期内的所有刷卡数据, 统计该乘客分别在各站点的刷卡次数, 再分别除以再所有站点刷卡次数的总和, 可以得到高频站点的吸引权重, 记作 A_i [3]。式 2 中 i 为高频站点的个数, A_i 为在一段时期内 IC 卡数据记录中该名乘客在第 i 个刷卡地点的刷卡次数, A_i 为频点的吸引权(频次)。

$$A_1 = \begin{cases} 1, & i = 1 \\ A_i \div \sum_{n=1}^i A_n, & i > 1 \end{cases} \quad (2)$$

站点下车吸引权：表示此站点对于所经线路的乘客在此站点下车的吸引力度。先求出该条线路在此站点上车人数的总和，再分别除以该线路此车次的总计上车人数。式 3 中 K_j 为本车次在第 j 个站点的上车人数， n 为站点个数。

$$A_2 = K_j \div \sum_{n=1}^j K_n, T_i \notin R \quad (3)$$

3.2. 下车站点算法

收集整理城市连续 7 天的 IC 卡数据和公交 GPS 数据，在已知公交站点和确定乘客上车站点的条件下，构建乘客出行链[4] [5]分析乘客的出行行为，对乘客下车站点进行判断。算法如下：(见图 4)。

Step 1: 汇总某乘客的所有出行记录，对照连续两次出行线路 L_i 与 L_{i+1} ，先判断这两条线路是否为同一条线路，然后根据 GPS 数据中的时间记录和上下行信息判断下车站点是否处在上车站点的下游。

Step 2: 如果 Step 1 的判断结论为该下车站点并非与上车站点在同一条线路且位于下游处，则获取下一条刷卡记录的 GPS 坐标 (x_{i+1}, y_{i+1}) ，计算其与上车刷卡线路 L_i 的欧式距离 O_{i+1} ，判断 O_{i+1} 是否小于 1 km (一般同一条线路相邻的两个公交站点间距离不超过 1 km)。

Step 3: 若判断连续两条线路 L_i 和 L_{i+1} 为同一条线路，且后一次刷卡记录位于前一次行程下车站点的下游，或者 O_{i+1} 小于 1 km，则认为乘客出行链连续，后一条刷卡记录的上车站点即为上一条刷卡记录的下车站点；

Step 4: 若以上判断均不成立，则认为该乘客的出行节发生断裂。此时则需获取该名乘客的出行规律，再次判断。查找乘客公交 IC 卡 ID 编号，根据编号提取该名乘客在连续 7 天内的所有公交出行刷卡记录，

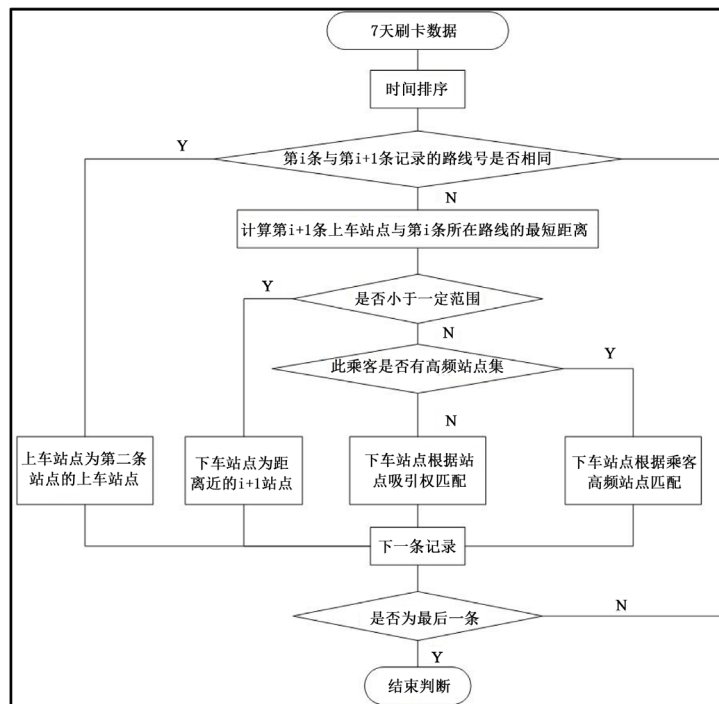


Figure 4. Flowchart about off site recognition algorithm

图 4. 下车站点识别算法流程图

记录汇总后计算其分别在每个上车站点的上车频次, 依此判断前次出行上车点的下游是否存在高频站点;

Step 5: 若上步判断结果得到下游存在高频站点, 则认为下游的这个高频站点就是该乘客此次出行的可能下车站点, 频次 A_{1i} 越高则乘客在此站下车的可能性越大;

Step 6: 若 Step 4 判断结果得到下游不存在高频站点, 此时我们利用该乘客此次出行所乘线路 L_i 在各站点的吸引权重判断乘客下车站点的可能性。提取线路 L_i 的线路编号 $line_ID$, 筛选出该乘客在对应编号线路上的全部刷卡数据, 结合站点位置信息, 计算乘客在各站点的下车吸引权 A_{2i} ;

Step 7: 按照 Step 6 乘客在下游站点下车吸引权的计算结果, 确定乘客此次出行的最可能下车站点。某站点的吸引权越大, 则乘客在此站下车的可能性就越大, 反之, 吸引权越小, 乘客在此站下车的可能性则越低。

通过上述算法进行乘客下车站点的识别。

4. 误差分析

挑选深圳市 M352 次公交的 15 个趟次, 站点客流情况如图 5。

分别进行上、下车站点推算 T 检验。检验结果如表 1、表 2。

其中, 变量 1: IC 站点上车客流量; 变量 2: 人工调研站点上车客流量。

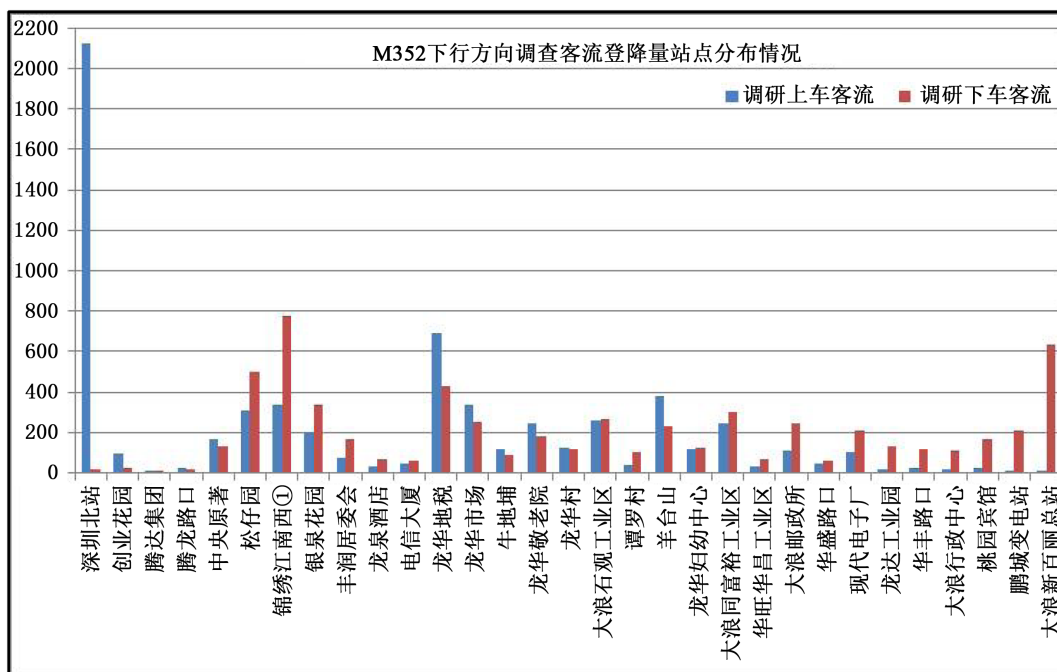


Figure 5. Passenger flow distribution about M352

图 5. M352 客流站点分布情况

Table 1. Get-on passenger flow treatment result

表 1. 上车客流处理结果表

P 取值	变量 1					
	IC 处理结果	调研结果的 1.04 倍	调研结果的 1.044 倍	调研结果的 1.05 倍	调研结果的 1.1 倍	
变量 2	调研结果	0.90255	0.91142	0.90278	0.88995	0.78728

其中, 变量 1: IC 站点上车客流量; 变量 2: 人工调研站点上车客流量。

Table 2. Get-off passenger flow treatment result
表 2. 下车客流处理结果表

P 取值	变量 1				
	IC 处理结果	调研结果的 1.06 倍	调研结果的 1.065 倍	调研结果的 1.07 倍	调研结果的 1.08 倍
	变量 2				
调研结果	0.79160	0.80783	0.79268	0.77769	0.74819

其中, 变量 1: IC 站点下车客流量; 变量 2: 人工调研站点下车客流量。

其中, 变量 1: IC 站点下车客流量; 变量 2: 人工调研站点下车客流量。

可以看出, 当 IC 卡站点客流量分别取调研结果扩样和 IC 处理结果时, 人工调查客流量与 IC 卡处理结果的 T 检验参数取值接近。因此, 可以大致认为, IC 处理结果的整体误差比例在 5%左右。

5. 结语

本文先对公交 IC 卡刷卡数据和 GPS 数据采用时间匹配的方法得到近似上车位置, 再按照修正时间差优化上车位置, 再依此为基础通过 DBSCAN 聚类分析的方法判断上车站点。最后根据出行链的思想, 给予乘客高频站点下车吸引权, 分析不同出行行为下可能的下车站点。最后结合深圳市的公交运行数据, 用 T 检验对算法进行误差检验。整个识别流程具有一定的实际意义和应用价值。

基金来源

桂林电子科技大学研究生教育创新计划(2016YJXC06)。

参考文献

- [1] 徐建闽, 熊文华, 游峰. 基于 GPS 和 IC 卡的单线公交 OD 生成方法[J]. 微计算机信息, 2008, 24(8): 221-222.
- [2] 陈锋, 刘剑锋. 基于 IC 卡数据的公交客流特征分析——以北京市为例[J]. 城市交通, 2016(1): 51-58, 64.
- [3] 李佳怡. 应用智能公交和路网数据的城市公交站点出行计算模型与评价[D]: [硕士学位论文]. 太原: 太原理工大学, 2017.
- [4] 尹安藤. 基于公交 GPS 和 IC 卡数据的公交 OD 推算[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2017.
- [5] 宋子杭. 基于公交 GPS 和 IC 卡数据的公交线网优化方法[D]: [硕士学位论文]. 北京: 北京交通大学, 2017.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2326-3431, 即可查询
2. 打开知网首页 <http://cnki.net/>
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>
 期刊邮箱: ojtt@hanspub.org