

文章编号: 1000-5641(2017)05-0201-12

智能交通刷卡记录中的公交站点恢复方法

王艺霖, 章志刚, 金澈清

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 随着城市公共交通网络的快速发展以及智能交通卡的普及, 智能交通卡中隐藏着越来越丰富的个人及群体移动行为信息. 但当前很多城市智能公交卡主要用于收费功能, 并未包含乘客确切的上下车时间及站点信息, 这给分析挖掘交通卡刷卡数据、提供基于精确位置的服务带来了阻碍. 本文针对上海市不含公交上下车站点的刷卡数据集, 借助于确定的地铁站点刷卡信息, 分析个人的整体刷卡历史记录, 提出一个基础的基于时空邻近性的恢复算法(STA, Space-Time Adjacency algorithm)和一个改进的基于历史的恢复算法(HTB, Historical Trip Based algorithm). 具体地, STA 算法根据刷卡记录线路的时空邻近关系进行恢复, 在此基础上, HTB 算法将刷卡记录集合根据时间和空间属性进行切分, 获得有明确出行意义的出行记录, 再利用历史记录集合, 提取乘坐线路以及频繁换乘线路, 根据线路间的空间关系生成线路带权候选站点列表, 再次进行站点恢复. 实验证明本文算法可以较好地缩小线路的候选上下车站点范围, 且时间效率较高.

关键词: 智能交通卡; 缺失数据; 刷卡数据挖掘; 站点推测

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3969/j.issn.1000-5641.2017.05.018

Individual station estimation from smart card transactions

WANG Yi-lin, ZHANG Zhi-gang, JIN Che-qing

(School of Data Science and Engineering, East China Normal University, Shanghai 200062, China)

Abstract: With the fast development of public transportation network and widespread use of smart card, more and more rich semantic information about human mobility behaviors are hidden in smart card transaction data. However, a great number of current smart cards are initially designed for charging and do not record any detailed information about where and when a passenger gets on or gets off a bus, which brings out great difficulties for analyzing, mining transaction data and providing more precise location-based services. This paper presents Space-Time Adjacency algorithm (STA) and Historical Trip Based algorithm (HTB) to estimate the bus station of each card's transaction records with the aid of integral historical data including complete subway transaction data. Specifically, STA does the initial reconstruction work according to the space-time proximity of adjacent

收稿日期: 2017-06-30

基金项目: 国家重点研发计划重点专项(973)(2016YFB1000905); 国家自然科学基金(61370101, 61532021, U1501252, U1401256, 61402180)

第一作者: 王艺霖, 女, 硕士研究生, 研究方向为基于位置的服务. E-mail: ylwang@stu.ecnu.edu.cn.

通信作者: 金澈清, 男, 教授, 博士生导师, 研究方向为基于位置的服务. E-mail: cqjin@sei.ecnu.edu.cn.

transaction records. Then HTB first cuts the collection of records to form trips that contain explicit trip purposes, then extracts taken lines and transfer lines using historical data, next generates candidate stations for each taken line, and finally uses them to recover the transaction records again. Experiments show that the proposed algorithms work well and narrow the range of candidate stations for bus lines, and have good time efficiency.

Key words: smart card; incomplete data; card mining; station estimation

0 引 言

随着世界人口的增加和城市人口比例的不断提高,设计、维持和促进可持续的城市公共交通模式变得非常重要.近年来,有越来越多的城市提供更加丰富的公共交通出行方式,促使更多乘客选择公交、地铁出行.与此同时,城市智能交通卡也在广泛普及,便捷的付费方式以及优惠的付费政策正吸引着越来越多的人采用智能交通卡出行.因此,每天都有数量巨大的智能交通刷卡数据在累积.大量的交易记录隐含着丰富的信息,它不但记录着一个人的公共交通出行基本信息,反映一个人的出行模式,也隐含着城市人群的移动模式和规律.目前有很多研究工作关注于利用智能交通卡数据分析挖掘乘客的移动模式^[1-2].文献[3]详细分析了利用智能交通卡数据进行人群移动行为分析的可行性,文献[4]从长期交通规划、公共交通服务调整、日常乘客乘坐需求分析等三个层面描述智能交通卡数据的应用.

尽管公共交通卡的广泛使用使得其隐含着关于乘客出行的丰富信息,但在有些城市中,公共交通卡的设计只为完成收费功能,卡中并未记录乘客出行的具体信息,如上下车站点、上下车时间^[5].例如,上海市公共交通卡可以在公交、地铁、出租车、轮渡等多种交通方式中使用,但只有地铁乘坐的上下车时间信息和站点信息被完整记录下来,出租车刷卡数据中只含刷卡时间而不含任何地理位置信息.由于上海的公交都采用一票制,在其交通卡信息存储的设计中就没有考虑存储上下车站点信息以及下车时间,只有乘客上车的时间被记录下来.数据集的不确定性和不完整性对分析和挖掘智能交通卡数据的研究工作产生了很大的阻碍.目前已有一些针对智能交通卡数据的恢复研究工作^[6-7],这些工作大多都对至少包含上车站点或下车站点其中之一的刷卡数据,利用“出行链”的思想进行数据恢复工作^[8].文献[5]针对部分公交线路中上下车站点信息均无的数据集,利用金钱、时间、空间维度的限制关系,以及占有一定比例的完整公交线路刷卡信息进行恢复,但其并没有利用乘客整体历史刷卡数据中的出行目的等隐藏含义以及乘客乘坐规律对站点进行恢复.以上研究工作均不适用于仅含有乘坐线路和上车时间的城市智能交通卡刷卡数据集的站点恢复工作.

为解决以上问题,本文提出基于历史出行记录的智能交通卡刷卡数据恢复方法.本文的主要贡献如下.

1) 考虑每条刷卡记录对于乘客的出行意义,提出了基于时间和空间的出行记录切分方法,以出行记录的维度进行分析和站点恢复工作;

2) 分析整张卡的所有历史出行记录,提取乘坐线路和频繁换乘线路,利用线路出现频次和线路间的空间位置关系,为线路站点设置权重,建立乘坐线路的候选站点列表,对刷卡记录中的站点进行再恢复;

3) 将所提出的方法应用于真实数据集中进行刷卡记录的恢复工作,分析证明了方法的合理性和有效性.

本文第1节介绍相关工作;第2节介绍数据集以及问题定义;第3节介绍基于时空邻近

性的刷卡数据恢复方法;第4节介绍基于出行记录切分和历史记录的站点推测方法;第5节进行实验结果展示与分析;第6节对工作进行总结和展望。

1 相关工作

本文工作主要与以下研究领域相关。一个研究领域为利用多种轨迹数据发现人群移动模式及规律,进行路线发现或推荐等;另一个研究领域为智能交通卡数据的恢复与挖掘工作。

随着各种轨迹数据,如出租车GPS数据、手机基站连接数据、公交刷卡数据等的不断积累,由历史出行数据中发现个人移动模式或群体移动模式或推荐路径引起了很多研究者的兴趣。一些研究工作表明,人们的移动模式有很强的规律性以及可预测性^[9]。在城市中,人们常遵循一定的时空出行规律,且主要活动在有限的几个固定地点附近,例如工作地和居住地,并在其中有规律地通行^[10]。文献[11]尝试发现目标的移动规律,包括在复杂的移动模式中找到移动周期,挖掘规律的移动行为等。在此基础上,衍生了很多有关人群移动数据的应用。文献[12]考虑人群移动的规律性和一致性,利用兴趣点签到数据、车载GPS数据、公交刷卡数据等预测人的移动位置。文献[13]利用出租车GPS数据构建轨迹数据库,记录出发及到达的时间地点,根据历史数据提供实时的路线费用及用时估计。文献[14]利用手机连接基站产生的GPS数据,发现停留区域,并获取有效移动轨迹,由历史轨迹数据发现热门线路。文献[15]利用海量出租车GPS历史数据,考虑时间、距离、油耗等因素,针对每位司机的出行偏好,筛选可参考的历史轨迹数据,提供实时路线推荐。海量历史轨迹数据隐含着丰富的信息,可以考虑个性化因素进行轨迹挖掘,提高推荐路线的质量。与GPS数据不同,公交刷卡数据记录一个人每天搭乘公共交通出行的历史轨迹,更能反映一些乘客常去的重要地点。通过对公交刷卡数据的分析,可以更好地了解城市公共交通的使用情况,提高服务质量。

与此同时,一些研究工作专注于智能交通卡刷卡数据的分析挖掘及补充和恢复工作。文献[16]总结了智能交通卡在城市研究中的应用,包括数据处理与上下车站点推测、公共交通系统的管理、城市空间结构的利用分析等几个方面。文献[17]利用北京市智能交通刷卡识别常用工作地、居住地以及频繁利用的上下班线路,研究城市上下班通勤模型。文献[18-19]分析公共交通乘坐行为,研究人们乘坐地铁或公交的可接受步行距离范围,发现影响步行距离最重要的因素是交通工具类型,而与出行目的、出行时间、乘客年龄等因素关系较小。

上述挖掘智能公交卡刷卡数据的工作常遇到公交刷卡数据信息不完整的问题。对于此问题,文献[8]首先提出了两条用于站点推测的重要假设:①大部分乘客当天最后一次出行的终点和当天第一次出行的起点相同;②大部分乘客上一次出行的终点与下一次出行的起点距离较近。多数恢复工作都利用了上述“出行链”的思想,主要针对上车站点或下车站点之一缺失的情况进行站点恢复工作^[6-8]。文献[20]对乘客刷卡时间进行聚类,与公交实时位置等其他数据来源进行匹配,辅助推测公交上下车站点;文献[21]建立公交站点吸引权系数概率模型,依据每个站点上下车乘客的数目及概率,推算乘客上下车站点,但其上下车站点概率的设置主要与站点热门程度相关,缺失针对一个人的整体历史记录进行站点推算的工作。目前只有文献[5]对上下车信息全无的公交刷卡记录进行恢复,但其仅考虑相邻刷卡记录而没有综合一个人的所有历史刷卡记录信息及乘客出行目的进行站点恢复工作。本文研究工作与其有以下几点不同:首先,文中提出了一种基于时空的刷卡记录切分方法,将刷卡记录组成有明确出行目的的出行记录;然后充分考虑整体出行记录中的隐含信息和线路间的空间关系,生成线路带权候站点列表,帮助确定上下车候选站点。

2 问题描述

在本节中, 主要进行数据准备及问题定义. 具体地, 2.1 节描述上海市公共交通刷卡数据集基本情况和公共交通网络的构建工作, 2.2 节给出基于以上数据集的问题定义.

2.1 数据描述

智能交通卡刷卡数据含上海 1 384 万张智能交通卡在 2015 年 4 月产生的 4.13 亿次刷卡数据, 刷卡数据类型包含公交、地铁、出租车、轮渡等. 其中公交专指公共汽车, 地铁指上海轨道交通, 含轨道交通 1 号线到 13 号线以及 16 号线, 共计 14 条线, 出租车指可以使用上海智能交通卡消费的城市出租车. 各种刷卡记录类型及数量如表 1 所示.

表 1 各类型刷卡数据数目统计

Tab. 1 Statistics of various transaction data types

类型	刷卡数	比例/%
地铁	2.48 亿	60.07
公交	1.55 亿	37.55
出租车	785 万	1.9
其他	198 万	0.48

每条刷卡记录包含以下属性: 卡号、日期、刷卡时间、交通工具类型以及线路名称. 其中地铁乘坐在进站和出站时都需要刷卡, 刷卡数据中包含了上下车站点及上下车时间; 公交只有上车时需要刷卡, 刷卡数据中仅包含线路名称和上车时间; 出租车刷卡数据中只包含下车时间; 还有小部分轮渡等不含地理位置信息的刷卡记录. 刷卡记录的具体格式如表 2 所示.

表 2 刷卡数据示例

Tab. 2 Charging records

卡号	日期	时间	类型	线路名称
2603642602	2014-04-06	11:45:31	公交	451路
3000706373	2014-04-18	11:22:26	地铁	2号线川沙
3000706373	2014-04-18	11:50:21	地铁	2号线金科路
2002816084	2014-04-26	20:55:00	出租车	无

城市公共交通网络由公交及地铁线路站点组成. 在刷卡数据集中, 共出现 1 344 条公交线路, 14 条地铁线路. 利用公共地图应用接口高德API, 查询刷卡数据集中出现的所有公交及地铁线路, 以及各线路站点的具体位置信息. 由于刷卡记录中有些线路名称有误, 以及少数公交线路运行调整, 有 89 条公交线路在地图应用接口中查询不到具体信息, 视为“未知线路”, 对应刷卡记录占总记录的 1.33%. 为保证出行轨迹的完整性, 保留“未知线路”的刷卡记录. 最终获取公交线路 1 255 条, 公交站点 12 740 个, 地铁线路 14 条, 地铁站点 360 个, 以构建公交和地铁线路信息列表.

2.2 问题定义

从上述数据集的描述中可以看出, 数据集中存储的有关乘客上下车地理位置描述的信息缺失严重, 公交线路的上下车站点全部都没有被记录. 数据的不确定性和不完整性给进一步挖掘人群移动模式带来了阻碍. 为更好地挖掘数据集中隐含的丰富信息, 提出公交站点恢复方法. 具体定义如下.

定义 1 (站点恢复) 给定一条公交刷卡记录, 考虑乘坐线路与邻近刷卡记录中乘坐线路交叉情况、线路间各站点间距、乘坐线路时间、乘客乘坐历史上下文等信息, 对该条刷卡记录中

的公交上/下车站点进行恢复.

3 基于刷卡记录时空邻近性的站点推测

一次公交刷卡记录意味着乘客利用公共交通进行了一次地理位置的移动. 相邻刷卡记录不但有着时间邻近性, 而且在乘客没有采用其他卡中无记录的交通工具出行的情况下, 上一次乘车的下车站点和下一次乘车的上车站点之间很可能具有空间邻近性, 因而可以采用基于时空邻近性的站点推测算法(STA, Space-Time Adjacency algorithm)进行站点推测.

刷卡数据中含有的刷卡类型有公交、地铁、出租车等, 有以下两种情况可以利用空间关系进行简单的站点推测. (1) 对于地铁-公交或公交-地铁的连续乘坐, 卡中可得知具体的地铁站点, 进而可以寻找公交线路中距离该地铁站最近的公交站点. (2) 对于公交-公交的连续乘坐, 在卡中无法获取到任何站点信息, 仅知道乘坐线路及上车时间. 这样的连续两次乘坐可能出现以下几种情况: (a) 因线路相同或出现“未知线路”, 未找到站点; (b) 线路距离较远, 未找到站点; (c) 线路重合较多, 可能的上下车站点较多; (d) 线路重合站点或距离相近站点较少, 可进行站点推测.

简单来说该算法寻找时间上相邻的两条线路的重合站点或者距离相近的站点作为恢复结果, 因而需要设置判定是否为邻近站点的距离阈值. 考虑到乘客步行速度限制、人群活动范围的有限性、以及当前获取换乘线路信息的便捷性, 人们更可能在相同站点或距离更近的站点进行下一次乘坐. 文献 [20] 中对人们步行情况的研究表明, 人们一天中的步行距离有限且大多小于 2 km. 文献 [21] 显示, 在加拿大蒙特利尔人们可接受的到公交站和地铁站的步行距离分别在 400 m 和 800 m 以内. 此外我们计算出上海市公交站平均间隔距离约为 700 m, 因此对于地铁-公交相邻乘坐和公交-公交相邻乘坐, 分别设置距离阈值 d_1 , d_2 , 默认 $d_1 = 1.5$ km, $d_2 = 1$ km.

具体过程如算法 1 所示. 首先对刷卡数据按照卡号进行分组, 使得同一张卡的记录分到同一组当中(line 1), 然后遍历数据集对每一张卡的所有记录按时间排序(line 3), 接下来对排好序的刷卡记录进行遍历, 考虑前后相邻的记录, 利用距离阈值寻找候选站点(line 5).

算法1 基于刷卡记录时空邻近性的站点推测算法STA

```

输入: 刷卡记录集合  $D$ , 距离阈值  $d_1, d_2$ 
输出: 乘客乘坐线路上下站点结果集  $ResultMap$ 
1   $PartitionMap \leftarrow \text{partition}(D)$ 
2  FOR each  $card$  in  $PartitionMap$  DO
3     $sortedList \leftarrow \text{sortList}(PartitionMap.get(card))$ 
4    FOR each  $record$  in  $sortedList$  DO
5       $ResultMap \leftarrow \text{findStation}(record_i, record_{i-1}, d_1, d_2)$ 

```

对刷卡记录数据集进行统计分析, 其邻近乘坐线路的空间邻近性具体情况如表 3 所示. 可以发现大多数的连续乘坐具有空间邻近性. 但同时发现公交-公交乘坐模式中, 34.30% 的连续乘坐是相同线路, 另外 20.50% 的连续乘坐线路之间存在超过 3 个较近的站点, 这给刷卡数据的推测和恢复工作带来了挑战.

表 3 线路连续乘坐情况统计

Tab. 3 Adjacent rides condition

乘坐模式	距离较远/%	同一线路/%	距离较近/%
公交-地铁	13.15	—	86.85
公交-公交	3.68	34.30	41.52 (相同站点数 ≤ 3), 20.50 (相同站点数 > 3)

4 基于历史出行记录的站点推测

通过以上的分析和统计,可以看出仅考虑连续刷卡记录的时空邻近性的站点推测方法有很多局限性,对于连续乘坐同一条线路或连续乘坐相同站点较多的记录不能得到很好的恢复,且没有考虑每条刷卡记录的潜在含义.本节介绍基于历史出行记录的站点推测算法(HTB, Historical Trip Based algorithm).结合文献[8]中有关“出行链”的思想,该算法有两个重要假设:(1)使用智能交通卡的乘客大多数有固定的居住地,所以各天的第一次出行记录大多由居住地附近出发,各天的最后一次出行记录也大多回到居住地.(2)前一天最后一次出行记录若未回到居住地,可能与第二天的第一次出行记录的起点有着空间邻近性.

基于历史出行记录的站点推测算法除进行算法1的站点推测处理外,还进行以下三个处理操作:出行记录划分;提取乘坐线路及频繁换乘线路;挖掘公交线路候选上下车站点,对数据进行再恢复.

4.1 出行记录划分

人们的一次出行,有具体的出行时间、出发地和目的地,对应智能交通卡中的一条或多条刷卡记录.刷卡记录中的一条线路或者可以使乘客从出发地直达目的地,或者是乘客为到达目的地而乘坐的线路之一.为更好地利用刷卡记录中的隐藏含义,提出一种基于时间和空间的记录切分方法,将刷卡记录组成出行记录.出行记录的具体定义如下.

定义2 (出行记录) 一条出行记录是由 n ($n \geq 1$)条刷卡记录按时间顺序构成的序列,且满足以下三个约束:(1)相邻刷卡记录根据其交通工具类型的不同,刷卡时间间隔小于特定的时间阈值;(2)连续两次地铁刷卡记录构成一次完整的地铁乘坐且包含在同一次出行记录中;(3)同一条线路的连续两次乘坐一定被包含在两次不同的出行记录中.

乘客在一次有明确目标的出行中会尽快完成乘坐及换乘,以抵达目的地.因而一张卡的刷卡记录集合中,大于一定时间阈值的相邻两次刷卡记录被认为属于两次不同的出行,应该被划分到两次出行记录中.一次出行中,乘客完成换乘和乘车两种行为,设置换乘时间阈值 T_1 和乘坐时间阈值 T_2 .对于一次出行记录内的公交-公交/公交-地铁的连续乘坐,阈值设为 $T_1 + T_2$;对于地铁-公交/出租车的连续乘坐,由于地铁刷卡发生在出站时,两次刷卡时间间隔仅包含换乘时间 T_1 .

出行记录切分的具体方法如算法2所示.给定一张卡的所有刷卡记录和时间阈值,遍历刷卡记录,根据切分规则进行切分.地铁数据包括进站和出站信息,因而必定成对出现,判断当前记录是否属于出站记录,决定是否对该记录进行处理(Line 4).然后使用设定的时间阈值参数,按照本节提出的阈值划分规则进行划分,将结果存入出行记录集合 L (Line 5-7).

算法2 出行记录切分算法

输入: 一张卡的所有刷卡记录集合 E , 换乘时间阈值 T_1 , 乘坐时间阈值 T_2

输出: 出行记录集合 L

```

1  初始化出行记录集合 $L$ 
2  初始化新的出行记录的起始位置pos为0
3  FOR  $i=1$  to  $|E|-1$  DO
4    IF ( $E[i]$ 不是卡中连续出现的第偶数条地铁刷卡记录)
5      IF ( $\text{cut}(E[i-1], E[i], T_1, T_2)$ ) /* 刷卡记录满足切分规则*/
6        从 $E[\text{pos}]$ 到 $E[i-1]$ 组成一条新的出行记录并加入 $L$ 
7         $\text{pos} \leftarrow i$ 
```

此外,一条出行记录在一天中所有出行记录中的相对位置以及出发时间与出行目的地有着较强的关联.例如一张卡在某天有两条出行记录,且第一条出发时间为早上,该条出行记录的起点更可能在居住地附近.若一张卡在某天仅有一条公交刷卡记录,且乘坐时间为晚上10:00,该次乘坐的下车站点更可能靠近居住地.分析一张卡的出行记录时间分布状况,利用出行记

录的出发时间辅助判断出行目的,有利于站点的推测工作.具体地,根据一天中的出行记录次数及其在一天中的时间段,将出行记录分为以下五种:START(一天中多条出行记录中的第一条记录)、END(一天中多条出行记录中的最后一条记录)、MID(一天中多条记录中除去标签为START和END的出行记录)、ONESTART(一天中唯一的出行记录且为由居住地附近出发)和ONEEND(一天中唯一的出行记录且为回到居住地附近).

4.2 乘坐线路及频繁换乘线路发现

虽然乘客的出行路线多种多样,但仍有相当比例的卡遵循着自己在时间和空间上的出行规律,其每天的出发地或到达地相对固定.因而可以利用乘坐线路间的站点位置关系,推测上下车站点.同时发现在一条出行记录中,往往包含多条刷卡记录,也就是乘客的一次出行需要多条公交线路的组合才能到达.而这种频繁的换乘行为恰恰说明换乘的公交线路及换乘的地铁线路没有距离其出发地较近的站点,因而可以利用这种信息为后续处理提供帮助.

定义3(乘坐线路) 乘坐线路包括标记为ONESTART或START的出行记录中的第一条乘坐线路,和标记为ONEEND或END的出行记录中的最后一条乘坐线路.

定义4(换乘线路) 换乘线路是指标记为ONESTART或START的出行记录中的第二条乘坐线路,和标记为ONEEND或END的出行记录中的倒数第二条乘坐线路.

算法3 乘客乘坐线路及频繁换乘线路提取

输入: 一张卡的所有出行记录集合 L , 频繁换乘线路频次阈值 $freq$

输出: 乘客乘坐线路及频次集合 $LineMap$, 频繁换乘线路及频次集合 $TransferMap$

```

1  初始化集合 $LineMap$ ,  $TransferMap$ 
2  FOR  $i=0$  to  $|L|-1$  DO
3    IF ( $L_i$ 的标签为ONESTART或START)
4       $LineMap.update(L_i.r_0.line)$ 
5      IF ( $L_i.r_0$ 类型为公交&& $L_i.r_1$ 类型为地铁或公交)
6         $TransferMap.update(L_i.r_1.line)$ 
7    ELSE IF ( $L_i$ 的标签为ONEEND或END)
8       $LineMap.update(L_i.r_{last}.line)$ 
9      IF ( $L_i.r_{last}$ 类型为公交&& $L_i.r_{last-1}$ 类型为地铁或公交)
10        $TransferMap.update(L_i.r_{last-1}.line)$ 
11  FOR each  $l$  in  $TransferMap$  DO
12    IF ( $l$ 的出现频次 $<freq$ )
13      将 $l$ 从 $TransferMap$ 中移除

```

在HTB算法中,主要对标签为START/ONESTART的出行记录的上车站点及标签为END/ONEEND的出行记录的下车站点进行再推测,缩小乘客的候选站点的范围.具体地,乘客乘坐线路及频繁换乘线路提取方法如算法3所示.给定出行记录集合和频繁换乘线路阈值,遍历出行记录.首先判断当前出行记录的标签类型,如果是ONESTART或START,将其第一条刷卡记录对应的线路加入到乘坐线路列表中,并更新其频次(Line 3-4),同时如果第一条刷卡记录为公交,则将第二条刷卡记录中出现的地铁站点或公交线路加入到换乘线路列表中,并更新其频次(Line 5-6).相对应地,对于标签为ONEEND和END的出行记录,处理过程相似,将出行记录中的最后一条和倒数第二条刷卡记录对应的线路分别加入到乘坐线路列表和换乘线路列表中并更新频次(Line 7-10).最后,使用设定的频繁换乘线路阈值参数 $freq$,对换乘线路列表进行过滤,删除频次过低的换乘线路(Line 11-13).

4.3 公交线路候选站点发现

算法3中提取的乘坐线路列表中可能含多条公交线路或多个地铁站点,它们出现的频次以及每两条线路间的空间邻近关系各不相同.这些线路之间拥有共同的公交站点或经过相同的区

域, 而公交上下车站点很可能在这些线路共同经过的区域内. 同时, 频繁换乘线路列表中一些站点也会相交在一片共同区域中, 显然此类站点不是目标站点, 根据这个辅助信息对候选站点进行筛选.

具体的候选站点挖掘过程如算法 4 所示. 给定乘坐线路列表 $LineMap$ 和频繁换乘线路列表 $TransferMap$, 考虑线路出现的频次以及线路间的站点邻近关系, 首先筛选 $LineMap$ 中的线路, 在 $LineMap$ 中删除频繁换乘线路列表中的地铁站点以及公交线路 (Line 2-4). 然后将 $LineMap$ 中的线路组成公交-公交线路对和公交-地铁线路对, 线路对的权重取为两线路在 $LineMap$ 中的频次之和. 遍历组成的线路对, 找出线路间相同或满足线路站点距离阈值 d_2 的站点对, 这些站点对将加入对应线路的候选站点列表, 两线路对产生的所有符合距离阈值的站点将平分该线路对的权重, 更新线路中候选站点的权重 (Line 5-9). 接下来删除结果列表中的频繁换乘站点, 对每一条线路, 选择权重最高的站点作为最可能的上下车站点, 若有几个权重最高且均相同的候选站点, 将其一起保留 (Line 10-11).

算法4 公交线路上下车候选站点生成算法

输入: 乘坐线路列表 $LineMap$ 频繁换乘线路列表 $TransferMap$,

输出: 换乘线路的候选站点列表 $CandidateMap$ <线路名称, Map <站点名称, 权重>>

```

1  初始化公交线路的候选站点列表  $CandidateMap$ 
2  FOR each  $l$  in  $LineMap$  DO
3    IF(线路  $l$  出现在  $TransferMap$  中)
4      将其从  $LineMap$  中删除
5  FOR each  $line_i$  in  $LineMap$  DO
6    FOR each  $line_j$  in  $LineMap$  DO
7      IF( $!line_i = line_j$ )
8        List<Station>  $stations = findCandidateStations(line_i, line_j)$ 
9        updateCandidateMap( $stations, CandidateMap$ )
10 FOR each  $lineCandidate$  in  $CandidateMap$  DO
11   removeStation( $CandidateMap, LineMap$ )

```

经过以上处理过程, 对于每一条非频繁换乘线路的公交乘坐线路, 都产生了一个候选站点列表, 存储着候选站点及其权重. 利用这个结果可以对算法1的结果中标签为 START 或 ONESTART 的出行记录的出发站点以及标签为 END 或 ONEEND 的出行记录的到达站点进行再恢复, 缩小线路的候选站点范围.

5 实 验

5.1 实验数据集

实验采用上海市政府数据服务网公开的城市智能刷卡数据集^[22], 数据集描述如第 2 节所示. 此外, 本文选取了 100 位志愿者的卡进行人工标注. 表 4 介绍了标注数据集中卡的出行记录数目分布, 表 5 介绍了卡的乘坐线路数目分布.

表 4 人工标注数据出行记录数目分布

Tab. 4 Distribution of cards' trip number

出行记录数目范围	百分比/%
1~5	24
6~10	21
11~15	20
16~20	9
> 20	26

表 5 人工标注数据乘坐线路数目分布

Tab. 5 Distribution of taken lines' number	
乘坐线路数目范围	百分比/%
1~2	22
3~4	32
5~6	20
> 6	26

5.2 实验环境及相关设置

实验在拥有 24 个节点的集群中完成, 操作系统为 Ubuntu 12.0.4. 每个节点搭载 6 核 Intel(R) Xeon(R) CPU E7-4809 v2 @1.90 GHz 的处理器, 内存共 50 GB. 所有实验使用 JAVA 代码实现, JDK 版本为 1.8.0. 算法 2 中换乘时间阈值 T_1 设为 30 分钟, 乘坐时间阈值 T_2 设为 30 分钟, 算法 3 中频繁换乘线路阈值参数 $freq$ 设为 3.

5.3 实验效果分析

整个数据集中, 有 20.53% 的卡只有地铁和出租车刷卡记录, 有 18.22% 的卡只含一条有效公交线路, 对于这两种卡进行过滤, 不进行站点推测工作, 以下实验仅针对需要恢复的数据进行操作. 本文实现了第 3 节提出的基于刷卡记录时空邻近性的站点推测方法, 以及第 4 节提出的基于历史出行记录的站点推测方法. 实验效果分析 HTB 算法对整体数据集的处理效果, 以及 STA 算法和 HTB 算法在标记数据集上的恢复结果.

(1) 整体数据集算法效果分析

由算法 2 划分所得的出行记录内部, 相邻记录间不具有空间邻近性的记录占比 1.71% (其中还包括与出租车的连续乘坐), 与表 3 中所有邻近刷卡记录间的空间邻近性统计情况相比, 可以发现所设时间阈值范围内, 同一出行记录内部确实具有更强的空间邻近性关系, 符合换乘的一般距离规律, 这也证明了记录切分方法的合理性. 所有出行记录的标签分布结果如表 6 所示. 标记为 MID 的出行记录仅占总出行记录的 14%, START 和 ONESTART 的出行记录与 END 和 ONEEND 的出行记录占主要部分且比例相近. 这说明大多数情况下, 一天中一张卡的出行记录不超过两条, 即一天中乘客利用公共交通卡的出行不超过两次.

表 6 出行记录的标签占比统计

Tab. 6 Ratio of various labels on trips	
出行记录标签	比例/%
MID	14
START	36
ONESTART	7.3
END	36.1
ONEEND	6.6

图 1 展示了经算法 3 处理后得到的乘坐线路数目以及从乘坐线路中筛选掉频繁换乘线路后的乘坐线路数目分布. 可以看出, 有很少的卡仅拥有一条乘坐线路(即在一个月內每天第一条和最后一条乘坐线路全部相同), 拥有 2-8 条乘坐线路的卡最为常见. 丰富的乘坐线路给利用线路间的空间关系来获取候选站点提供了可能. 当从初步提取的乘坐线路中删除掉频繁换乘线路后, 整体分布趋势为卡的乘坐线路减少. 被筛选掉的频繁换乘线路可以防止换乘站点在算法 4 中权重设置过高, 进而提高推测的准确性.

图 2 展示了 STA 算法和 HTB 算法中标记为 START/ONESTART 的出行记录的公交上车候选站点数目以及标记为 END/ONEEND 的出行记录的公交下车候选站点数目累积分布对比.

无任何公交候选站点的出行记录的数目由 46.42% 降到 5.24%。图中可以看出, HTB 算法明显减少了候选上车站点的数目。HTB 算法中, 仅有一个候选站点的出行记录数目达到 STA 算法的 2.8 倍。推测结果中, 仍然会有一小部分出行记录的候选站点数目较多, 出现此种推测结果的原因可能是乘客刷卡数据集中乘坐线路有限, 或这些线路的重合站点较多, 1 个月的刷卡数据集中提供信息较少, 不利于充分推测线路站点。

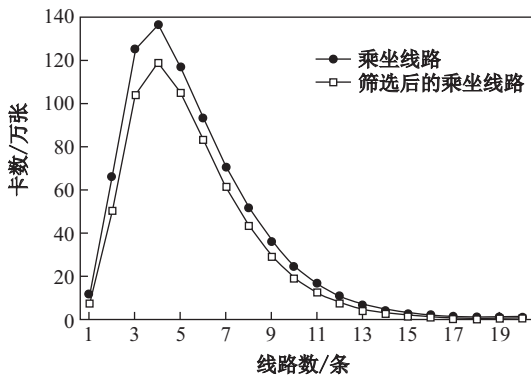


图1 乘客乘坐线路数目分布

Fig. 1 Distribution of taken lines

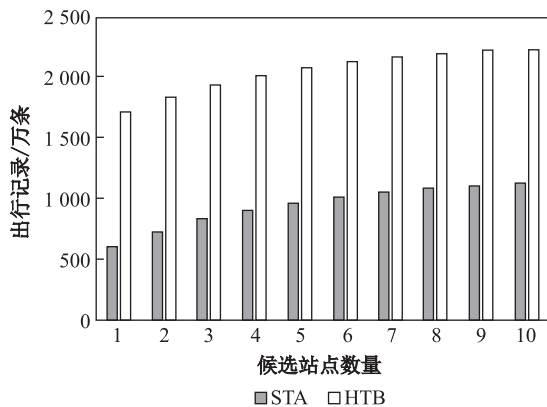


图2 候选站点数目分布

Fig. 2 Distribution of candidate stations' number

(2) 标注数据集算法性能分析

评价算法准确度具体从准确率(Precision)、召回率(Recall)、 F_1 值(F_1 -measure)这三个方面进行考量。若算法推测的站点与人工标注的站点相差在两站之内, 则认为有效推测出了该条刷卡记录的一个上/下车站点。假设算法找出的站点个数为 P , 其中正确找出的站点个数为 Q , 人工标注出的站点个数为 R , 于是有 $\text{Precision} = Q/P$, $\text{Recall} = Q/R$, $F_1\text{-measure} = 2 \times PR / (P + R)$ 。利用标注数据对 STA 算法和 HTB 算法进行性能分析, 结果如表 7 所示。

表 7 算法性能对比

Tab. 7 Performance of comparison

算法	准确率/%	召回率/%	F_1 值
STA	48.6	34.7	0.41
HTB	78.9	85.1	0.82

由表 7 可以看出, 相比 STA 算法, HTB 算法的准确率和召回率均较高, 说明 HTB 算法的

有效性. 同时 STA 方法准确率比召回率高, 而 HTB 算法的召回率比准确率高. 这是因为 STA 算法仅考虑相邻刷卡记录进行站点推测, 对于连续两条相同公交线路乘坐的情况和相邻刷卡记录为出租车的情况, STA 算法不做站点推测, 导致 P 值较小, 准确率相对召回率有所提升. 而由于 HTB 算法的策略是利用历史出行记录尽可能对所有乘坐线路的站点进行恢复, 这导致有更大的可能性使得每条线路都产生候选站点列表, 进而使得 P 值较大, 例如对那些拥有较多个候选站点的线路依然会进行恢复, 而不是放弃恢复, 造成召回率高于准确率.

5.4 算法运行性能分析

设置测试数据集大小依次为整个数据集的25%、50%、75%、100%, 分别运行两种推测算法. 图3展示了两种站点推测方法在不同大小数据集下的运行时间, 可以看出算法整体运行时间与数据集大小成线性关系. HTB算法运行时间约为 STA 算法的 3.5 倍, 但是从之前的分析来看, HTB算法的准确度远远高于 STA 算法, 因而时间开销是可接受的.

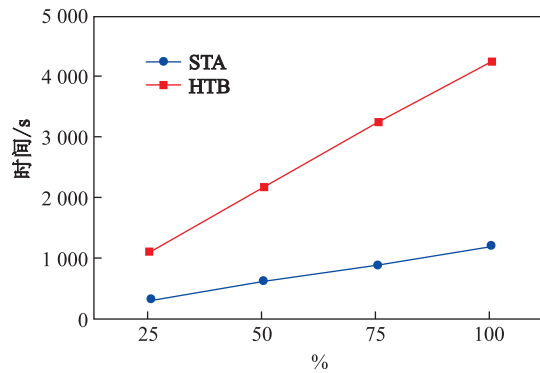


图3 HTB和STA运行时间

Fig.3 Running time of HTB and STA

图4则是 HTB 算法各个步骤的时间消耗状况, 可以发现算法1(STA)、算法2和算法3的时间消耗相对较小, 算法4占用大部分的运行时间. 这是因为 HTB 算法在出行记录划分算法中的时间与刷卡记录数目 n 成线性关系, 时间复杂度为 $O(n)$; 在寻找乘坐线路及频繁换乘线路中与出行记录数目 m 成线性关系, 时间复杂度为 $O(m)$; 而最后的生成候选站点算法需要首先生成线路对, 然后在线路对中寻找候选站点并计算权重, 时间复杂度较高.

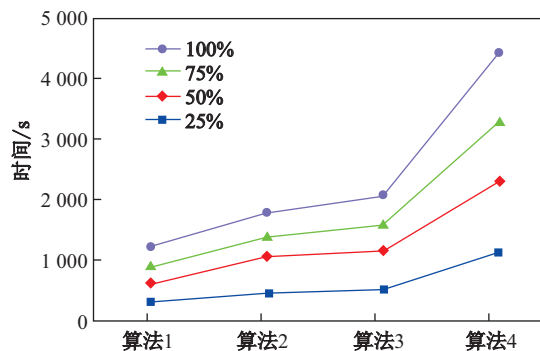


图4 HTB算法各步骤运行时间

Fig.4 Running time of each procedure in HTB

6 总 结

本文针对公交上下车站点缺失的城市智能交通卡刷卡数据,对公交站点进行推测,提出了基于时空邻近性的站点推测方法(STA)以及基于历史出行记录的站点推测方法(HTB)。STA 算法只考虑用邻近刷卡记录的乘坐线路之间的空间关系进行恢复,而 HTB 还构建了出行记录,结合每张卡的历史出行记录对站点进行细粒度的恢复。实验表明 HTB 算法比 STA 算法大大减少真实刷卡记录中公交候选上下车站点的推测范围,提高了推测站点的准确性。

[参 考 文 献]

- [1] LATHIA N, CAPRA L. How smart is your smartcard? Measuring travel behaviours, perceptions, and incentives[C]// Proceedings of the 13th International Conference on Ubiquitous Computing. ACM, 2011: 291-300.
- [2] LATHIA N, FROELICH J, CAPRA L. Mining public transport usage for personalised intelligent transport systems[C]// 2010 IEEE 10th International Conference on Data Mining. IEEE, 2010: 887-892.
- [3] BAGCHI M, WHITE P R. The potential of public transport smart card data[J]. Transport Policy, 2005, 12(5): 464-474.
- [4] PELLETIER M P, TRÉPANIÉ M, MORENCY C. Smart card data use in public transit: A literature review[J]. Transportation Research Part C Emerging Technologies, 2011, 19(4): 557-568.
- [5] ZHANG F, YUAN N J, WANG Y, et al. Reconstructing individual mobility from smart card transactions: A collaborative space alignment approach[J]. Knowledge and Information Systems, 2015, 44(2): 299-323.
- [6] TRÉPANIÉ M, TRANCHANT N, CHAPLEAU R. Individual trip destination estimation in a transit smart card automated fare collection system[J]. Journal of Intelligent Transportation Systems Technology Planning & Operations, 2007, 11(1): 1-14.
- [7] WANG W, ATTANUCCI J P, WILSON N H M. Bus passenger origin-destination estimation and related analyses using automated data collection systems[J]. Journal of Public Transportation, 2010, 14(4): 131-150.
- [8] BARRY J, NEWHOUSER R, RAHBEE A, et al. Origin and destination estimation in New York City with automated fare system data[J]. Transportation Research Record, 2002, 1817: 183-187.
- [9] SONG C, QU Z, BLUMM N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327: 1018-1021.
- [10] GIANNOTTI F, NANNI M, PEDRESCHI D, et al. Unveiling the complexity of human mobility by querying and mining massive trajectory data[J]. The VLDB Journal, 2011, 20(5): 695-719.
- [11] LI Z, DING B, HAN J, et al. Mining periodic behaviors for moving objects[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 1099-1108.
- [12] WANG Y, YUAN N J, LIAN D, et al. Regularity and conformity: Location prediction using heterogeneous mobility data[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1275-1284.
- [13] BALAN R K, NGUYEN K X, JIANG L. Real-time trip information service for a large taxi fleet[C]// Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. ACM, 2011: 99-112.
- [14] DASH M, KOO K K, HOLLECZEK T, et al. From mobile phone data to transport network—gaining insight about human mobility[C]// IEEE International Conference on Mobile Data Management. IEEE, 2015: 243-250.
- [15] DAI J, YANG B, GUO C, et al. Personalized route recommendation using big trajectory data[C]// IEEE 31st International Conference on Data Engineering. IEEE, 2015: 543-554.
- [16] 龙瀛, 孙立君, 陶遂. 基于公共交通智能卡数据的城市研究综述[J]. 城市规划学刊, 2015, 3: 70-77.
- [17] LONG Y, THILL J C. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing[J]. Computers Environment & Urban Systems, 2015, 53: 19-35.
- [18] EL-GENEIDY A, GRIMSRUD M, WASFI R, et al. New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas[J]. Transportation, 2014, 41(1): 193-210.
- [19] DANIELS R, MULLEY C. Explaining walking distance to public transport: The dominance of public transport supply[J]. Journal of Transport & Land Use, 2011, 6(2): 5-20.
- [20] CUI A. Bus passenger origin-destination matrix estimation using automated data collection systems[D]. Cambridge, MA: Massachusetts Institute of Technology, 2006.
- [21] 胡继华, 邓俊, 黄泽. 结合出行链的公交 IC 卡乘客下车站点判断概率模型[J]. 交通运输系统工程与信息, 2014, 14(2): 62-67.
- [22] 上海市数据服务网. [DB/OL]. [2017-05-20]. <http://www.datashanghai.gov.cn>.

(责任编辑: 林 磊)