计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(4):32−38 [doi: 10.15888/j.cnki.csa.006863]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于出租车数据的载客热点与打车热点研究①

陈丽璐, 聂文惠

(江苏大学 计算机科学与通信工程学院, 镇江 212013)
通讯作者: 陈丽璐, E-mail: 1006885594@qq.com

摘　要: 面对城市出租车高空载率和乘客打车难问题, 本文针对出租车司机端和乘客端分别进行载客热点和打车热点的分析研究, 提出了一种基于 DBSCAN 算法的数据处理模型. 利用这个模型对北京市 182 辆出租车的 GPS 轨迹数据进行处理, 提高了数据精度；对于不同的受众, 采用 K-means 算法对数据进行聚类分析, 得到相关热点. 实验表明, 划分目标用户进行各热点的推荐不仅可以有效地为出租车司机提供高概率的载客热点, 乘客打车难问题也有了一种可行的解决方法.

关键词: 出租车数据; GPS 轨迹数据; 载客热点; 打车热点

引用格式: 陈丽璐,聂文惠.基于出租车数据的载客热点与打车热点研究.计算机系统应用,2019,28(4):32–38. http://www.c-s-a.org.cn/1003-3254/6863.html

Research on Passenger Hotspots and Taxi Hotspots Based on Taxi Data

CHEN Li-Lu, NIE Wen-Hui

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Faced with the problem of high no-load rate of urban taxi and taxi difficulty of passengers, this study analyzes the passenger hotspots for taxi drivers and taxi hotspots for passengers, and proposes a data processing model based on DBSCAN algorithm. Using this model, the GPS trajectory data of 182 taxis in Beijing are processed, and the data precision is improved. For different audiences, K-means algorithm is used to cluster the data and get the relevant hotspots. Experiments show that the proposed method can not only effectively provide taxi drivers with high probability of passenger hotspots, but also provide a feasible solution to the problem of taxi difficulty of passengers.

Key words: taxi data; GPS trajectory data; passenger hotspots; taxi hotspots

随着人们生活水平不断提高, 城市车辆数量急剧上升, 城市交通问题随之日益突出, 人民日常出行驾驶车辆常常会遇到交通堵塞情况, 尤其是在北京、上海、广州等大城市, 居民如果想快速、方便地出行, 特别是在暴雨、寒冬等恶劣天气条件下, 出租车无疑是众多公共交通工具中的最佳选择之一. 出租车数量的不断增长也使出租车行业面临了越来越多的问题, 如空驶率高、分布不均衡、供不应求等等, 这些问题也一定程度上导致了居民打车难的问题.

近年来, 随着全球定位系统 GPS (Global Position System) 技术的飞速发展和智能定位终端的广泛应用, 基于位置的信息服务得到了飞速发展. 城市出租车上基本都安装了 GPS 传感器, 能够记录当前车辆的位置、采集时间等信息. 很多国内外学者基于出租车轨迹信息展开了各项研究, 例如: 路径规划[1,2]、基于位置的社交网络[3]、智能交通系统[4]和城市计算[5−7]等. 文献[8]通过对首尔的出租车 GPS 数据进行多维分析, 指出可以运用数据分析手段指导出出租车运营公司进行合

① 收稿时间: 2018-10-23; 修改时间: 2018-11-12, 2018-11-19; 采用时间: 2018-11-23; csa 在线出版时间: 2019-03-28

理调度.文献[9]应用 K-means 聚类模型分析出租车载客的热点区域, 应用遗传算法进行出租车的应急调度; 文献[10]将 K-means 算法和具有噪声的基于密度的聚类算法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN) 在轨迹数据研究方面进行了对比, K-means 算法在处理大数据时效率较高, 但对于噪声点和孤立点数据较为敏感; DBSCAN 算法不需要进行预处理亦可在带有噪声点的数据中发现任意形状的聚类, 但对输入参数的敏感性较高.

本文提出了一种服务于出租车司机和乘客的载客热点与打车热点的研究方法. 主要研究步骤为: 首先对轨迹数据进行预处理, 形成基础数据集; 接着对基础数据集采用速度与时间结合算法进行出租车停留点和出租车载客点的提取; 然后对出租车停留点进行 DBSCAN 聚类得到核心停留点; 最后对核心停留点和出租车载客点分别进行 K-means 聚类分析得到推荐热点.

1 轨迹点预处理

由于轨迹数据量巨大, 并且对于城市交通来说, 频繁拥堵状态下缓慢行驶车辆的 GPS(全球定位系统) 定位信息易产生冗余定位点与噪声, 例如建筑物会对信号产生阻隔, 就会存在一些信号缺失的现象, 因此需对轨迹数据进行预处理.

1.1 相关定义

轨迹 G 通常由一系列的包含时间和空间信息的点组成, 表示为$G = P1,P2,\cdots,Pn$, 其中, n 为一条轨迹的点的总数; $Pi(1\le i\le n)$ 为轨迹点, 且 $Pi=(X, Y, T)$, X 和 Y 分别为第 i 个轨迹点的经度和纬度坐标, T 为时间戳, 有 $Pi.T<Pi+1.T$.

定义 1. 出租车停留点 Sp (Stay Point of Taxi)

根据实际生活常识, 通常出租车在载客状态时, 其驾驶的平均速度会超过一定的阈值, 而在没有载客时, 为了寻找乘客, 其行驶速度通常低于某个值. 因此设定出租车在载客时的平均驾驶速度高于阈值 Vs, 未载客时速度低于该阈值.

在出租车 GPS 轨迹点中存在这样一些点, 此类点的驾驶速度 V 小于给定的速度阈值 Vs, 并且以小于该速度阈值的速度连续行驶一段时间 T, 而 T 大于给定的时间阈值 Ts, 同时此类点所在的移动范围小于给定的距离阈值 Ds, 那么我们就认为此类点是出租车停留点.

定义 2. 核心停留点 Sc (central stay point)

给定距离阈值 ε、最小密度数 m, 给定一条轨迹中的轨迹点 P, 记 N 为 P 的 ε 距离邻域内同属于该轨迹点的点的集合, 如 $|N|\ge m$, 则称 P 为核心停留点.

定义 3. 出租车载客点 St (Pick-up Point of Taxi)

出租车搭载乘客的地点, 如图 1 所示, 当行程 1 中的平均速度小于速度阈值, 行程 2 中的速度超过一定速度阈值, 且行程 2 的行驶距离超过一定的距离阈值, 则定义停留点 S1 是一个历史载客点. 当行程 3 的速度大于一定的速度阈值时, 行程 3 作为出租车载客行驶的一部分行程.

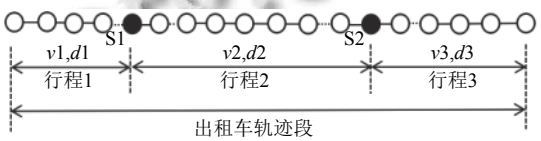

图 1 出租车载客点示意图

出租车停留点 Sp 是由出租车停留或徘徊所产生的点的集合, 核心停留点 Sc 是对出租车停留点进行 DBSCAN 算法聚类得到的点的集合, 出租车载客点 St 是根据出租车的行驶速度和时间进行判断得出的点的集合. 三者的关系如图 2 所示.

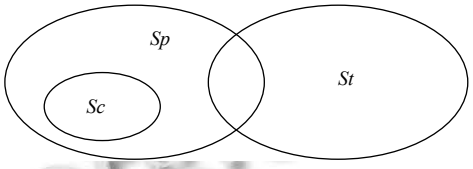

图 2 定义点关系图

1.2 轨迹点处理

根据以上的定义, 在提取出租车停留点和出租车载客点之前, 首先需要对轨迹数据中的漂移数据进行处理. 本文采用 Douglas-Peucker(曲线数据压缩) 算法进行轨迹数据的漂移处理. 算法步骤如下: 首先, 提取三个位置相邻的数据点分别为点 A, 点 B 和点 C, 如图 3, 根据三角形面积公式, 三角形 ABC 的面积为:

$$s = \sqrt{p(p-a)(p-b)(p-c)} \tag{1}$$

其次, 判断点 B 到前后两点 A 与 C 的距离是否满足公式:

$$h = \frac{2s}{c} < derror \tag{2}$$

其中, $derror$ 是指允许偏移的最大值, 一般由地图精度、GPS 接收机精度和道路宽度之和计算得出. 比较 h 与 $derror$ 之间的大小, 如果小于 $derror$, 则直线 AC 段作

为曲线 AC 的近似, 该段曲线处理完毕; 如果 h 大于 $derror$, 则点 B 不可忽略, 将曲线 AC 近似为直线 AB 和 BC 两段.

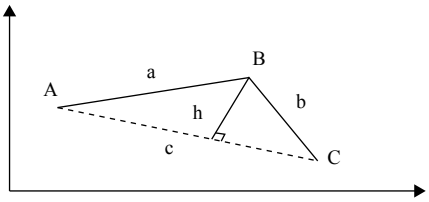

图 3 Douglas-Peucker (DP) 线路优化示意图

接着对轨迹数据进行噪声处理, 根据定义 1 中出租车停留点的含义, 在出租车 GPS 轨迹点中, 停留点并不仅仅代表速度为零的轨迹点, “停”只是 GPS 轨迹点的一种状态, 因此本文通过计算和分析出租车 GPS 数据, 提取出租车停留点, 进而获取相关热点.

YE[11]、KARLI[12]等人将移动对象活动分为四种模式: 内空间模式 (inside)、伴随模式 (along)、环绕模式 (around)、驻留模式 (stop). 如果出租车 GPS 轨迹数据符合环绕模式或驻留模式中其中一种模式, 我们就将该 GPS 轨迹点视为出租车停留点. 如图 4 所示, 图中点均为出租车 GPS 轨迹点, 圆圈内的点表示出租车停留点, Sp1 中没有路线规律的 GPS 轨迹点是由驻留模式产生得到出租车停留点, 而 Sp2 中有一定路线规律的 GPS 轨迹点是由驻留模式产生得到出租车停留点.

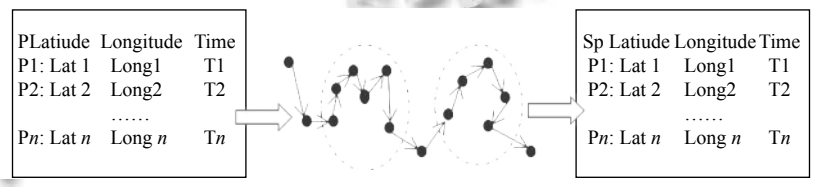

图 4 停留点提取示意图

驻留模式下的停留点均符合出租车停留点的定义, 但也可能存在一种特殊情况需要分析处理, 即驻留模式下由交通状况产生的停留点. 比如等待交通信号灯或者道路堵塞等状况, 将这种情况下产生的停留点视为噪声. 进一步分析研究得知, 等待交通信号灯和交通堵塞大部分情况是发生在交通路口, 为了处理这些噪声, 参考谭川豫等人[13]对移动对象轨迹的研究, 定义如果在某处有 50% 以上的出租车轨迹发生了方向转变, 并且转变角度超过 30%, 将该处视为交通路口, 并对该处周围 50m 范围内的所有出租车停留点视为噪声, 予以删除.

2 热点处理

本文从出租车司机端和乘客端两方面考虑, 对于出租车司机端进行载客热点的推荐, 乘客端进行打车热点的推荐, 尽可能降低出租车司机的空载率、减少乘客的等待时间. 为了达到更好地推荐效果, 对轨迹点根据时间这一因素划分并进行聚类处理. 对出租车停留点采用基于密度的聚类算法得到核心停留点, 对核心停留点和出租车载客点采用基于划分的聚类算法进行聚类.

2.1 划分思想

2.1.1 用户划分

受众目标的考虑不仅包括出租车司机还包括乘客, 对于出租车司机我们进行载客热点的推荐, 而对于乘客我们进行打车热点的推荐.

出租车司机的载客热点由两部分组成. 一部分是由轨迹数据处理得到的出租车停留点进行 DBSCAN 初步聚类得到的核心停留点; 另一部分是提取得到的出租车载客点.

对于乘客更多的是需要考虑出租车可能停留的地点, 所以对于打车热点的聚类研究是基于出租车停留点的相关数据基础上的, 即对核心停留点直接进行聚类研究.

2.1.2 时间划分

众所周知, 人类的活动受到空间、时间及其他限制条件的影响, 为了研究这些限制条件的影响, 瑞典地理学家赫格斯特兰德提出了时间地理学的概念[14], 其认为人类在时空中的活动受到三类限制: (1) 能力限制, 指生理或物理因素对人类果冻产生的限制, 如我们必须分配一定的时间来满足吃饭、睡觉等生理需求; (2) 权利限制, 指因活动场所常由不同的人或单位控制

而对他人产生的限制, 如, 某商场的营业时间是上午九点至晚上十点, 一般顾客只有在此时间段内可以进入这个商场的空间; (3) 结对限制, 指我们在特定的地点从事某项活动 (如到上班地点工作、到商场购物), 或必须和其他人共同从事某项活动 (例如开会), 因此要在时间和空间上彼此相互协调配合. 人们应在这些限制条件下合理的安排时间和空间, 以满足我们的生理、经济、社交等各方面的需求.

出租车司机的驾车行为和乘客打车行为均属于人类活动, 同样受上述三类限制, 如果协调好司机与乘客之间的行为, 司机在相应的时段到乘客相对集中的地点拉客人, 乘客在相应时段到司机经常路过的地点打车, 这样不仅司机能快速地找到乘客降低空载几率, 也方便了乘客打车. 图 5 选取了 8 点至 10 点和 14 点至 18 点两个时间段部分停留点可视化效果图, 该图验证了不同时段停留点的分布情况不同, 说明人群的集中地在各个时段内的分布情况不同. 车辆的载客情况会随时间的变化而变化, 划分时间能更好地进行热点推荐有一定的数据支持.

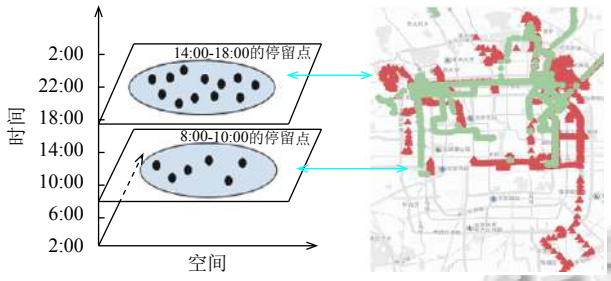

图 5 8:00–10:00(圆形点) 和 14:00–18:00(三角点) 两个时段部分停留点的情况

2.2 算法思想

本文处理轨迹数据的算法思路如下: (1) 对初始轨迹数据进行预处理, 预处理包括轨迹数据的漂移和噪声处理; (2) 对基础数据集 (即经过预处理后的数据集合) 结合时间和速度进行出租车停留点和出租车载客点的提取; (3) 对提取的出租车停留点进行 DBSCAN 算法进行聚类得到核心停留点; (4) 对核心停留点和出租车载客点进行 K-means 算法的聚类得到载客热点和打车热点. 具体过程如图 6 所示.

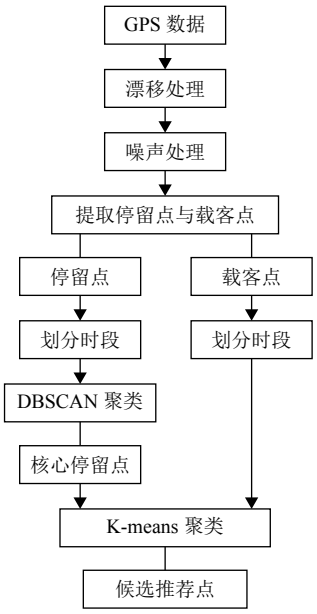

图 6 算法流程图

DBSCAN 算法是通过不断地搜索临近点, 使该点周围的密度逐渐增加, 以寻找到一个区域内所查找点密度大的地方. 出租车停留点是轨迹数据中符合环绕模式或驻留模式的轨迹点, 环绕模式或驻留模式的轨迹点大多聚集在一个区域内. 这些轨迹点所围绕的区域即是需要找的出租车停留区域, 故对这些停留点进行基于密度的聚类可以得到代表该区域的停留点, 即本文定义的核心停留点. DBSCAN 算法的相关定义如下:

(1) 密度直达: 给定距离阈值 ε、最小密度数 m, 则满足下述两个条件时, 点 P 是从点 Q 出发直接密度可达: P 属于 Q 的 ε 距离范围内, 记为 $P \in N$; P 为核心停靠点, 即 $|N| \geq m$.

(2) 密度可达: 给定自轨迹序列 $P1, P2, \cdots, Pn$, $P1=Q$, $Pn=P$, 对于任意 $i(1 \leq i \leq n-1)$, $Pi+1$ 是从 Pi 关于 ε 和 m 直接密度可达的, 则称点 P 是从 Q 关于 ε 和 m 密度可达的, 反之不一定成立.

(3) 密度相连: 若一条轨迹中存在一个点 O, 使得点 P 和 Q 是从关于 ε 和 m 密度可达的, 则称 P 和 Q 关于 ε 和 m 密度相连.

采用 DBSCAN 算法处理出租车停留点能快速且有效地处理噪声点和发现任意形状的空间聚类, 与 K-means 算法相比不需要输入要划分的聚类个数.

对于核心停留点和出租车载客点的聚类采用 K-means 算法进行聚类, K-means 算法的主要思想是通过

迭代的过程把数据集划分为不同的类, 从而使生成的每个类的内部数据紧凑, 类与类之间独立. K-means 算法处理大数据集具有良好的可伸缩性和高效性, 其缺点是需设定聚类簇的数量, 利用这一点可以实现对聚类中心的质量的控制.

3 实验结果与分析

本文的实验数据选取了北京市 7 个月的轨迹数据作为一个实验样本进行实验, 对上述方法进行验证与研究.

3.1 实验环境

实验计算机配置: CPUCore(TM)3.40 GHz, 内存 16 GB, 显存 8 GB; 操作系统: Windows 7; 使用软件: MATLAB, Pycharm, MySQL, SQLyog.

3.2 评价标准

为了评价本文获取的热点的准确率, 采用精度 (*Precision*, 即查准率) 和召回率 (*Recall*, 即查全率) 对相关热点进行评定[15]. 查准率是指正确识别的相似重复记录与实际的相似重复记录的比值, 查准率越高, 表明提取出来的相关热点精度越高. 召回率是各热点被正确识别的百分率, 召回率越高, 表明该方法识别各热点的能力越强. 查准率和召回率定义如下:

$$precision = \frac{tp}{tp + fp} \times 100\% \quad (3)$$

$$recall = \frac{tp}{tp + fn} \times 100\% \quad (4)$$

其中, *tp* 表示正确识别的热点数, *fp* 表示错误识别的热点数, *fn* 表示未识别的热点数.

文献[16]中认为在兴趣点半径 50 米范围内的出租车载客点都是正确的, 借鉴文献中判断出租车载客点的正确性的方法进行测试. 本文把各时段的出租车载客热点作为测试点, 地图上的兴趣点 (火车站、超市、公园、写字楼等) 作为已知点, 用测试点和已知点进行比较, 从而判断测试点的正确性.

3.3 实验分析

本文实验数据为微软公开的轨迹数据, 其中包含纬度、经度、海拔、时间等信息. 将数据集按时间进行时段划分, 总结处理后数据的时间分布规律, 将数据点按时间点分布所占比例进行分类, 如图 7 所示, 各时间段出租车的载客情况各有不同, 低谷期在凌晨到次日早高峰之间.

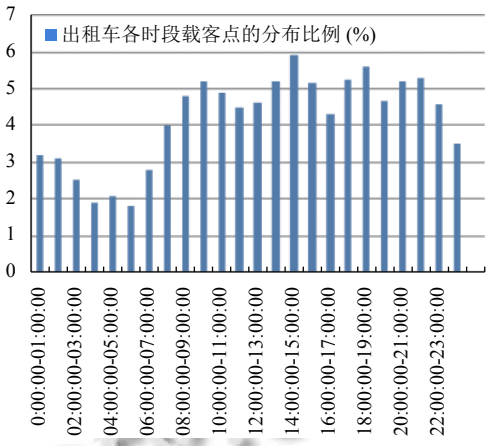

图 7 不同时段 GPS 数据点的分布比例

实验数据是每 5 秒采集一次, 对停留点的定义是符合环绕模式和驻留模式的点, 这两种模式会产生多个位置相近但不重复的点, 对于这些点进行 DBSCAN 聚类可以提高停留点的质量. 但对于在路口附近 50 米范围的点予以删除, 如图 8 所示, 小圆是经过 DBSCAN 聚类形成的簇, 而三角形的点和大圆中的点可能是拥堵或交通信号灯产生的点, 为避免影响停留点质量, 做删除处理. 停留点的优化一定程度上也提高了聚类的时间效率. 表 1 为部分停留点聚类结果.

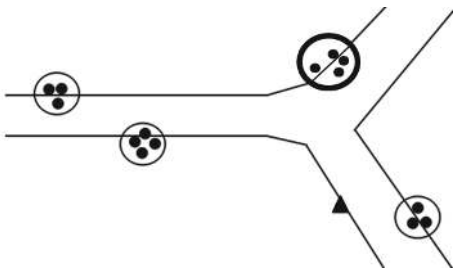

图 8 停留点聚类处理示意图

表 1 停留点聚类效果对比

样本	初始停留点个数	聚类处理后停留点个数
数据集 1	256	78
数据集 2	468	105
数据集 3	872	302

将本文获取出租车载客热点的方法 TDKC 分别与文献[17]中的层次聚类 (MSRA) 获取出租车载客点方法、文献[18]中的时空聚类 (STA) 获取出租车载客热点的方法进行比较, 结果如图 9 和与图 10 所示, 可以看出本文的方法获取的出租车热点在缓冲半径为 20 米时的精度和召回率分别达到了 85.9% 和 82%, 而当缓冲

半径为 50 米时, 本文方法的精度和召回率达到了 95.7% 和 95.5%, 较优于上述两种聚类方法, 95% 以上的精度和召回率说明了推荐的热点具有较高的推荐价值, 提高了推荐热点的准确性. 鉴于乘客的打车热点包含于出租车载客热点之内, 因此打车热点的准确性也有所提高.

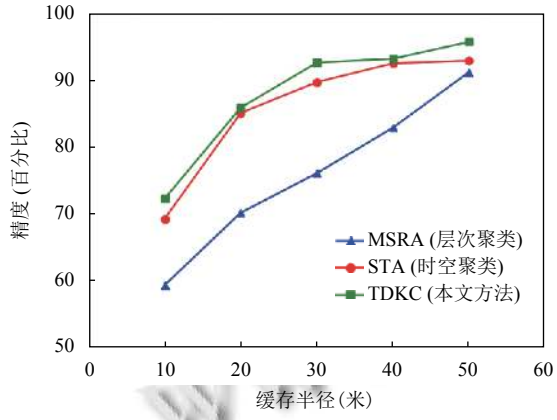

图 9 热点精度随缓存半径变化图

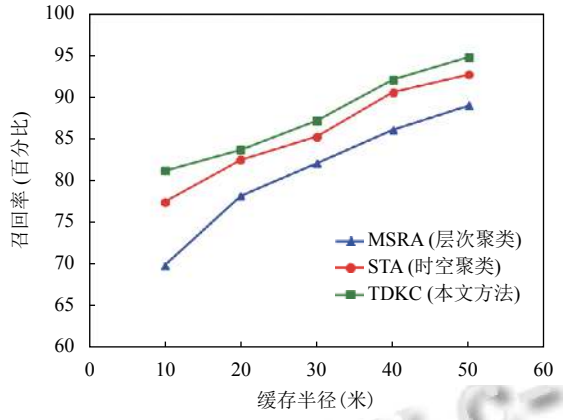

图 10 热点召回率对缓存半径变化图

4 结论

为降低出租车司机的空载率和减少乘客的等待时间, 本文改进了面向出租车司机的载客热点和面向乘客的打车热点的模型和算法. 在对 GPS 轨迹点的处理上, 对提取出来的出租车停留点采用 DBSCAN 算法进行聚类, 提高停留点的准确度; 由于时间因素对出租车司机和乘客的影响较大, 本文在对停留点和载客点聚类分析时, 运用划分思想, 将数据集按时间进行划分, 在各个分段中进行数据的聚类分析. 最后通过实验验证, 结合 DBSCAN 算法对停留点进行核心停留点的提取, 提高了停留点的准确性, 同时也提高了聚类分析的时间效率. 考虑时间因素对停留点和载客点进行分析在查准率和召回率方面都有所提高.

受实验环境的制约, 本文仅对部分数据进行了处理和分析, 后期将进一步对其他数据处理分析, 对各个候选热点根据推荐公式进行热点的推荐.

参考文献

1 Ziebart BD, Maas AL, Dey AK, *et al*. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. Proceedings of the 10th International Conference on Ubiquitous Computing. Seoul, South Korea. 2008. 322–331.

2 Yuan J, Zheng Y, Xie X, *et al*. T-drive: Enhancing driving directions with taxi drivers' intelligence. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 220–232. [doi: 10.1109/TKDE.2011.200]

3 连德福. 基于位置社交网络的数据挖掘[博士学位论文]. 合肥: 中国科学技术大学, 2014.

4 Yuan W, Deng P, Table T, *et al*. An unlicensed taxi identification model based on big data analysis. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(6): 1703–1713. [doi: 10.1109/TITS.2015.2498180]

5 Zheng Y, Liu Y C, Yuan J, *et al*. Urban computing with taxicabs. Proceeding of the 13th International Conference on Ubiquitous Computing. Beijing, China. 2011. 89–98.

6 Pan G, Qi GD, Wu ZH, *et al*. Land-use classification using taxi GPS traces. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1): 113–123. [doi: 10.1109/TITS.2012.2209201]

7 Zheng Y, Capra L, Wolfson O, *et al*. Urban computing: Concepts, methodologies, and application. ACM Transactions on Intelligent Systens and Technology, 2014, 5(3): 38.

8 Atmaji FTD, Sig KY. Mining the GPS big data to optimize the taxi dispatching management. Proceedings of the 4th International Conference on Information and Communication Technology. Bandung, Indonesia. 2016. 1–4.

9 温雅静. 基于热点载客区域的出租车应急调度方案研究[硕士学位论文]. 北京: 北京交通大学, 2014.

10 张致宁. 基于 K-means 和 DBSCAN 的轨迹数据挖掘研究. 中国战略新兴产业, 2017, (44): 113–114.

11 Ye Y, Zheng Y, Chen YK, *et al*. Mining individual life pattern based on location history. Proceedings of the Tenth International Conference on Mobile Data Management:

Systems, Services and Middleware. Taipei, China. 2009. 1–10.

12 Karli S, Saygin Y. Mining periodic patterns in spatio-temporal sequences at different time granularities. Intelligent Data Analysis, 2009, 13(2): 301–335. [doi: 10.3233/IDA-2009-0368]

13 谭川豫. 移动对象轨迹分析技术研究[硕士学位论文]. 长沙: 国防科技大学, 2010.

14 Hägerstrand T. Reflections on "What about People in Regional Science? ". Papers of the Regional Science Association, 1989, 66(1): 1–6. [doi: 10.1007/BF01954291]

15 成海霞. 基于 Android 的出租车载客热点推荐系统[硕士学位论文]. 湘潭: 湖南科技大学, 2016.

16 Yuan J, Zheng Y, Zhang LH, *et al*. Where to find my next passenger. Proceedings of the 13th International Conference on Ubiquitous Computing. Beijing, China. 2011. 109–118.

17 姬波, 叶阳东, 肖煜. 基于信息瓶颈方法的出租车空载聚集区聚类算法. 小型微型计算机系统, 2013, 34(9): 2139–2143. [doi: 10.3969/j.issn.1000-1220.2013.09.035]

18 张明月. 基于出租车轨迹的载客点与热点区域推荐[硕士学位论文]. 湘潭: 湖南科技大学, 2013.