Computer Science and Application 计算机科学与应用, 2018, 8(9), 1482-1489
Published Online September 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.89161

Hans汉斯

Extraction and Analysis of Hotspot Region of Parallel Taxi Trajectory Based on Spark

Xueli Li[1], Yong Sheng[2], Xiaoji Lan[1]

[1]School of Architectural and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi
[2]Shanghai Digital Intelligence System Technology Co., Ltd., Shanghai
Email: 1911375358@qq.com

Received: Sep. 10th, 2018; accepted: Sep. 23rd, 2018; published: Sep. 30th, 2018

Abstract

The taxi GPS trajectory data can mine wealthy residents travel law information, but for the increasing number of data, there are new requirements have been put forward about the accuracy and efficiency of data mining. This paper takes Chengdu taxi GPS trajectory data as the research object. First, the distortion of the original data and the redundant field should be deleted, and partial time data should be filtered, then the map should be matched; finally using the spark Big Data processing platform, it realized K-means| |, divided into working days and rest days to analyze and get the hot spot area of Chengdu residents and its space-time distribution characteristics. Finally, compared the performance of the K-means and K-means| |, the result showed that K-means| | had superiority in accuracy and time efficiency compared with the single machine.

Keywords

GPS Trajectory Data, Map Matching, Hotspot Area, Travel Hotspot

基于Spark的并行化出租车轨迹热点区域提取与分析

李雪丽[1]，盛　勇[2]，兰小机[1]

[1]江西理工大学建筑与测绘工程学院，江西　赣州
[2]上海数慧系统技术有限公司，上海
Email: 1911375358@qq.com

收稿日期：2018年9月10日；录用日期：2018年9月23日；发布日期：2018年9月30日

文章引用: 李雪丽, 盛勇, 兰小机. 基于 Spark 的并行化出租车轨迹热点区域提取与分析[J]. 计算机科学与应用, 2018, 8(9): 1482-1489. DOI: 10.12677/csa.2018.89161

摘　要

从出租车GPS轨迹数据中可挖掘出丰富的居民出行规律信息，但数据量的不断增加，对数据挖掘的准确性和效率提出了新的要求。本文以成都市出租车GPS轨迹数据为研究对象，首先对原始数据进行失真数据剔除、多余字段删除和部分时段数据过滤三方面的预处理，其次进行地图匹配，最后利用Spark大数据处理平台，实现K-Means||算法，分为工作日和休息日的不同时段进行挖掘分析，得到成都市居民出行热点区域及其时空分布特征，并将单机K-Means算法和K-Means||算法的性能进行对比分析，结果表明：相比于单机，K-Means||算法在准确性和时间效率上具有优越性。

关键词

GPS轨迹数据，地图匹配，热点区域，出行热点

1. 引言

随着城市中出租车数量的不断增多，GPS 卫星定位技术的不断发展与普及，装有车载 GPS 的出租车在行驶过程中产生了大量的出租车轨迹数据，记录了车辆的位置、时间、方位和速度等信息，通过挖掘出租车 GPS 轨迹数据可用于分析车辆移动轨迹特征、预测交通流、改善交通服务[1] [2] [3]，对城市交通管理、道路规划具有重要意义。

城市热点区域通常是人流量大、商业发达、经济水平发展高的中心地区，利用出租车轨迹数据提取城市热点区域的方法主要有根据数据场势值阈值法探测轨迹点的聚集模式进行提取[4]，基于高斯定律的轨迹挖掘方法[5]，将轨迹转换为网格序列进行聚类[6] [7]。但由于出租车 GPS 轨迹数据数量庞大且分布状态多样，因此对轨迹数据挖掘的方法提出了新的要求，需要研究高效的分布式并行轨迹数据挖掘算法[8] [9] [10]。借鉴传统研究方法之后，结合现在流行的 Spark 大数据处理平台，优化 K-Means 聚类算法[11] [12] [13] [14] [15]，将成都市出租车 GPS 轨迹数据进行研究，挖掘分析工作日休息日不同时段的数据，得到居民出行热点区域及其分布特点，并将单机 K-Means 算法和 K-Means||算法的时间效率进行对比，结果表明后者在处理大数据量的时间效率上有较好的效果。

2. 数据预处理

本文选取 2014 年 8 月 3~4 日成都市出租车 GPS 轨迹数据为实验对象，原始数据中包含出租车编号、经纬度、载客状态(1 代表载客，0 代表空车)、速度、方位角、时间戳七个字段，但由于受到噪声、通信故障和传感器硬件故障等外界因素的干扰，可能存在噪声数据、数据缺失和数据失真等现象，并不能直接用于处理和分析，因此需要从失真数据剔除、多余字段删除和部分时段数据过滤三方面对数据进行预处理。

2.1. 失真数据剔除

失真数据主要包括以下情况：① 轨迹数据经纬度超出成都市范围；② 源数据中存在信息不完整的

字段；③ 载客状态异常，存在除了 0 (空车)、1 (载客)外其他的数值，或该字段一直为 0 或 1，对以上异常数据均应剔除。

2.2. 多余字段删除

原始数据中速度和方位角对本文研究没有作用，故对这两个字段做删除处理，仅保留出租车编号、时间戳、经纬度、载客状态五个字段。

2.3. 部分时段数据过滤

00:00:00~05:59:59 时间段出租车基本处于停运状态，该时间段的轨迹数据对于提取居民出行高峰时段和挖掘分析城市热点区域没有参考价值，因此删除这段时间的轨迹数据。

出租车上下车位置的确定根据轨迹数据的“载客状态”字段来确定，上车点为邻近两点字段值由 0 变为 1，下车点则相反。居民日出行总量为上下车总次数的平均值，经统计得到工作日和休息日出行量随时间波动的时间序列如图所示，从图 1 中可看出，除 17:00~19:00 外，工作日居民出行量整体比休息日出行量高，工作日出行早高峰为 8:30~10:00，相比通勤早高峰推迟了半个小时，主要是由于出租车为中短途出行，为避免堵车会错开出勤高峰期。11:00~13:00 为午餐和休息时间，出行人数较少，进入一个小低谷，之后逐渐上升并保持稳定。18:00~19:00 为避开通勤晚高峰再次出现一个低谷，之后由于部分晚间公共交通停运，出租车需求增大，出行量逐渐增长并形成一个峰值，21:00 之后又开始渐渐降低。休息日相比于工作日而言，不受通勤行为影响，整体走势相对平缓，波动变化不大，17:00~19:00 也并未出现低谷。

3. 地图匹配

从 OpenStreetMap 上获得成都市基础地图数据，根据道路网的拓扑关系提取原始车辆位置周边的可能行车路段，对比出租车 GPS 轨迹数据与原始车辆位置附近的可能行车路段，根据相似度函数将形状相似度最高的可能行车路段作为最终的匹配结果，确定可能行车路段的函数定义如下：

$$S = q\sum_{i=0}^{n} D_i \tag{1}$$

$$D_i = \|T_i - C_i\| = \sqrt{(x_{Ti} - x_{Ci})^2 + (y_{Ti} - y_{Ci})^2} \tag{2}$$

式中，q 为加权系数，Di 为待匹配轨迹点与投影后的轨迹点的 2 范式，n 为待匹配轨迹点个数，Ci(x_Ci,y_Ci)与 Ti(x_Ti,y_Ti)分别为轨迹点的原始坐标和轨迹点 i 到路段的投影坐标。很显然，S 值越小，可能行车路段与 GPS 轨迹的相似程度越高，因此选择相似度函数最小的可能行车路段作为车辆的最终行驶路段。以成都市府青路到中环路为例，对比出租车 GPS 轨迹数据与基础地图数据进行匹配前后的效果差异，如图 2 所示。

4. 基本原理与方法

利用出租车 GPS 轨迹数据进行城市热点区域挖掘时，传统 K-Means 算法采取随机方式选择初始中心点，难以保证最终结果的可靠性，因此基于 Spark 平台，引入并行化 K-Means++算法：从样本数据中任意选择初始中心点 O，计算其余各点与初始中心点的最近距离 Dmin 并将其保存在集合 G 中，对 Dmin 求和得到 Sum，取任意一值 P (0 < P < Sum)，对 G 集合中的样本循环操作 P- = Dmin，直到 P < 0，该样本点就作为下一个初始中心点；重复以上步骤，直到 K 个初始中心点被选出来，并将得到的 K 个初始中心点带入 K-Means 算法进行迭代运算。计算与初始中心点的最近距离引用公式(3)，其中，SqDist 为样本点 point 与中心的点 center 的距离，为聚类中心点的位置坐标，为样本点的位置坐标。

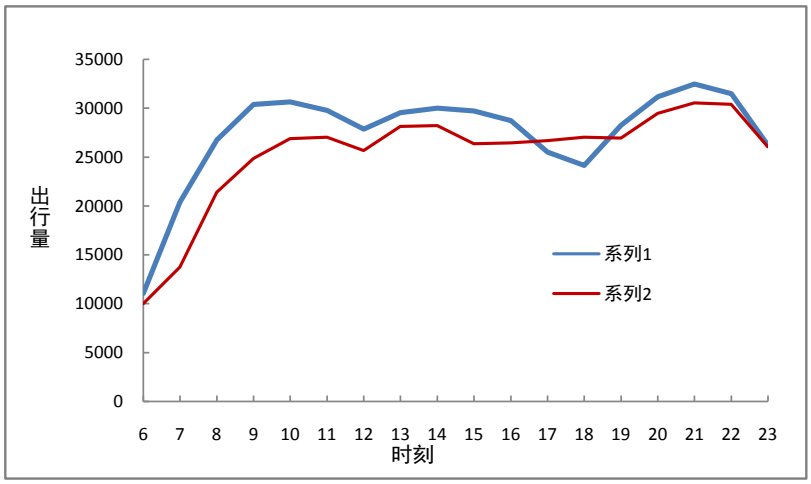
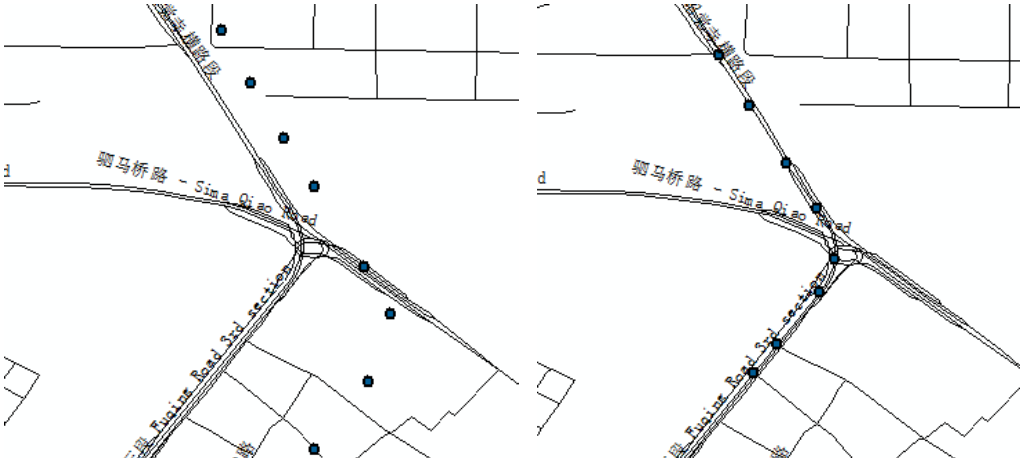


Figure 1. Comparison of travel volume of residents in each period
图 1. 各时段居民出行量对比



(a) (b)

Figure 2. The comparison of before and after map matching. (a) Before the matched map; (b) After the matched map
图 2. 地图匹配前后对比。(a) 地图匹配前；(b) 地图匹配后

$$\begin{aligned}\mathrm{SqDist} &= \left(\sqrt{a_1^2+b_1^2}-\sqrt{a_2^2+b_2^2}\right)^2 \\ &= a_1^2+b_1^2+a_2^2+b_2^2-2\sqrt{\left(a_1^2+b_1^2\right)\left(a_2^2+b_2^2\right)}\end{aligned} \tag{3}$$

不难发现，SqDist 小于等于欧式距离，且 SqDist 只需计算聚类中心点和样本点的 L2 范数，因此不再采用欧氏距离进行距离的计算。通过使用 K-Means++算法可获得最优初始中心点保证聚类结果的良好性，并在 Spark 平台上实现并行化，取名为 K-Means||算法，算法具体参数见表 1。

5. 实验与分析

实验平台采用 Spark on Yarn 模式进行搭建，由 3 台虚拟机和 2 台实体机组成，设置 1 个 Master 节点和 4 个 Slave 节点，Master 节点内存为 4G，Slave 节点内存为 2G，操作系统版本为 Centos7，JDK 版本为 jdk1.8.0。平台搭建完成并进行测试后，开始进行城市热点提取。

Table 1. K-Means|| detailed parameters
表 1. K-Means||参数详解

参数	参数含义
k	初始聚类中心点的个数
maxiterations	算法最大迭代次数
initializationMode	初始化方式，有两种：① 随机数选择初始中心点 ② 使用 K-Means||选择最优初始中心点
initializationSteps	K-Means||算法的运行步数
epsilon	K-Means||算法收敛的距离阈值
initialModel	设置初始中心点的可选参数，如果提供该参数，则表示 K-Means||算法仅被运行一次

5.1. 城市热点提取

对经过预处理的数据根据本文介绍的算法进行 K 值选择作为城市居民出行的热点区域数量，并通过聚类得到 K 个聚类中心点的位置以及每种类别轨迹点的个数。居民出行的密集程度的衡量通过计算热度值来评估(式(4))，热度值越大的地区，代表居民的出行的密集程度越高，该地也就需要较多的出租车。

$$dh_i = \frac{n_i}{m}, \tag{4}$$

其中，代区域编号为 i 的热度值，为该区域内的下车轨迹点数量，为下车轨迹点总数。对热度值进行等级划分，大于 0.1 设定为高热度，0.05~0.1 为中热度，小于 0.05 为低热度。以 8 月 4 日为例，早中晚三个时间段热点区域数量(K 值)分别为 11，10，12，早高峰期间热点区域中心位置、热点区域轨迹点数量、热度值和热度等级如表 2 所示。

将聚类结果可视化，分别得到成都市 8 月 4 日工作日期间早中晚三个时间段的车租车出行的居民热点区域分布情况，如图 3 所示。

由图 3 可发现，工作日早高峰(9:00~11:00)居民出行热点主要分布在工业基地和产业园一带，如少城视井文创产业园区、凉水井工业园、成都广告创业产业园，而景点和休闲娱乐场所未能形成明显的热点区域；工作日午高峰期间(13:00~15:00)，居民出行热点主要分布在商业中心和休闲娱乐场所，如天府广场、远东百货与铂金城购物广场，各大工业基地和产业园热点区域的数量和热度则明显下降；工作日晚高峰(21:00~23:00)，居民出行热点主要分布在城市居民居住聚集区，如九里提、李家沱、双林路片区，数量较多且分散，工业基地和产业园的密集程度急剧下降，而商业中心和休闲娱乐场所的居民聚集程度依然很高。工作日热点区域早中晚三个时间段的分布状况的变化符合人们上午外出工作，晚上回家休息的通勤特征。

对于休息日而言，居民出行没有明显的早高峰，全天分为午晚两个高峰时间段，热点区域数量(K 值)为 10，13，分布情况由图 4 可得，居民在休息日出行较晚，午高峰时间段为 13:00~15:00，热点区域主要分布在商业中心、休闲娱乐场所以及旅游景区，各大工业基地和产业园居民出行的聚集程度很低；晚高峰(21:00~23:00)居民出行热点区域分布在城市居民居住聚集区和商业中心，居民出行的聚集区域数量多且聚集程度高，在工业基地和产业园基本无聚集情况。

5.2. K-Means||算法性能分析

选取 8 月 4 日的 491453 个下车轨迹点和 8 月 3~4 日的 983074 个下车轨迹点使用单机 K-Means 算法和在 Spark 平台上使用 K-Means||算法进行 5 次聚类，聚类时间消耗如图 5 所示。由图可知，单机的运算速度要明显落后于集群，且随着集群的节点数量不断增多，运算速度明显加快。

Table 2. Distribution of early peak hotspots on August 4
表 2. 8 月 4 日早高峰热点区域分布情况

区域编号	所处位置	轨迹点数量	热度值	等级
0	104.029088, 30.699054	11,430	0.18667	高热度
1	104.082235, 30.65286	5987	0.097777	中热度
2	104.133713, 30.630555	9857	0.160981	高热度
3	103.957494, 30.573985	1798	0.029364	低热度
4	104.04544, 30.646584	4154	0.067841	中热度
5	104.120058, 30.700357	6925	0.113096	高热度
6	104.05694, 30.572859	3146	0.051379	中热度
7	104.071161, 30.681833	3682	0.060133	中热度
8	103.968098, 30.694703	2440	0.039849	低热度
9	104.002812, 30.646205	6130	0.100113	高热度
10	104.074672, 30.614757	5682	0.092796	中热度

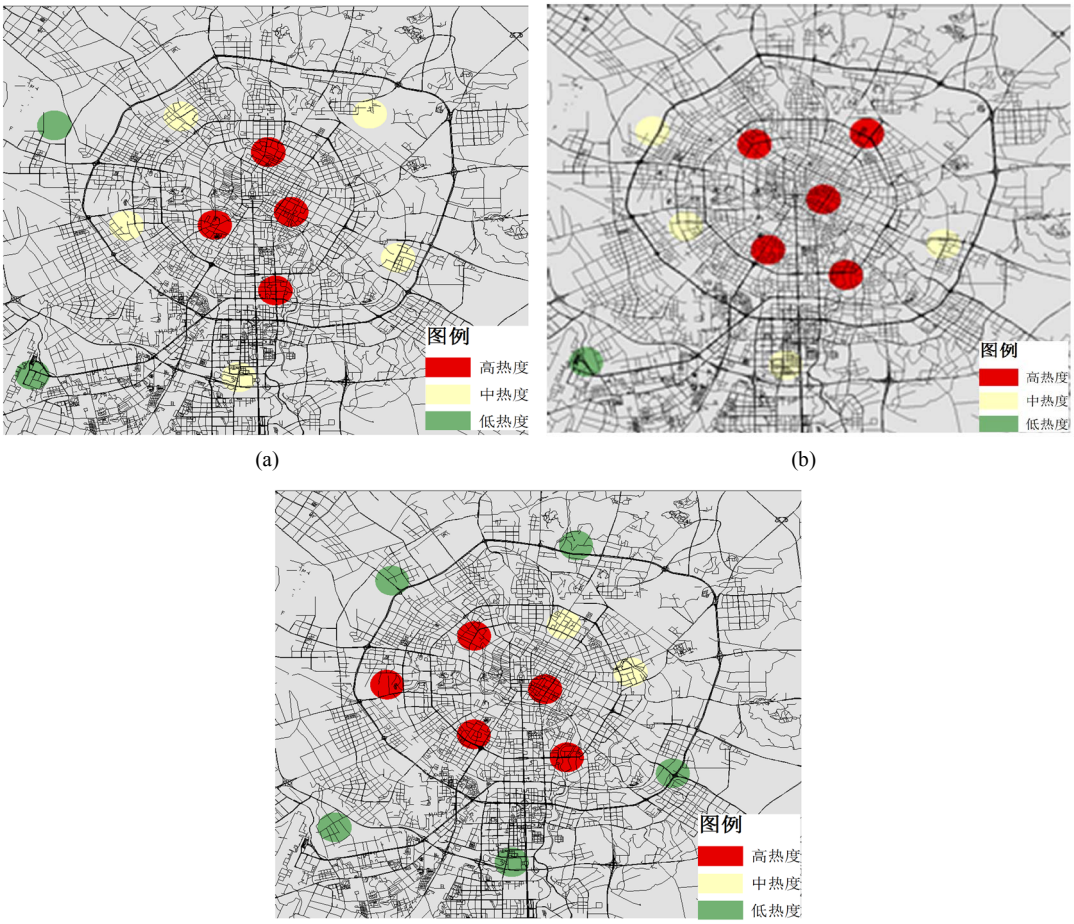
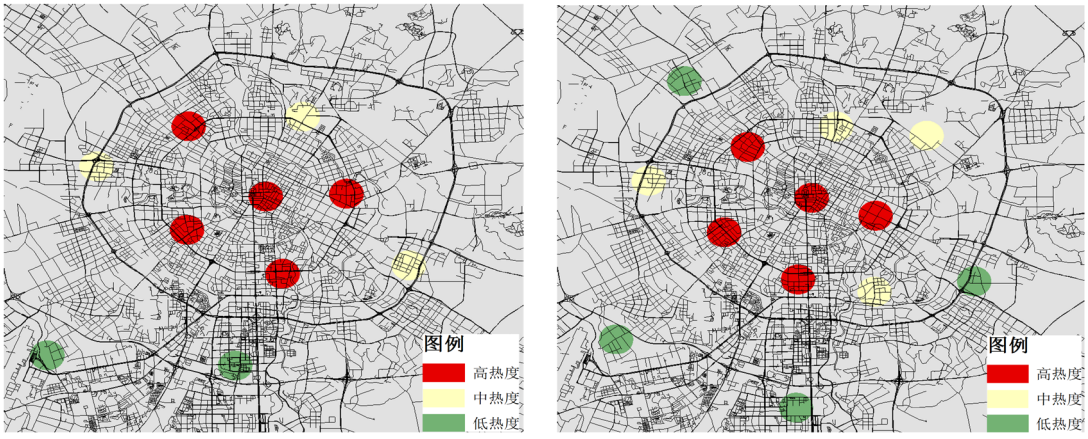


Figure 3. Distribution of hotspots during peak hours on workday. (a) Early peak distribution; (b) Midday peak distribution; (c) Late peak distribution
图 3. 工作日各高峰时段热点区域分布情况。(a) 早高峰分布情况；(b) 午高峰分布情况；(c) 晚高峰分布情况



(a) (b)

Figure 4. Distribution of hotspots during peak hours on weekend. (a) Midday peak distribution; (b) Late peak distribution

图 4. 休息日各高峰时段热点区域分布情况。(a) 午高峰分布情况；(b) 晚高峰分布情况

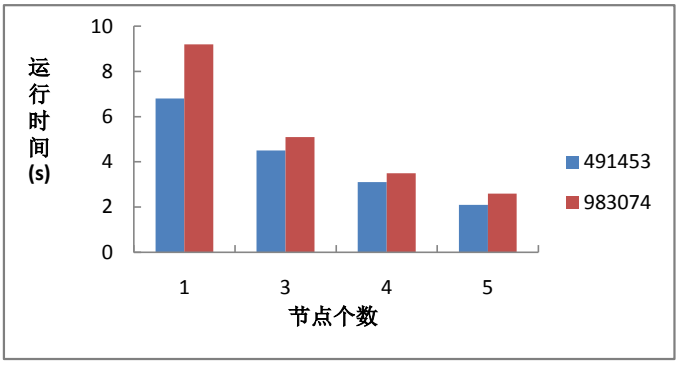


Figure 5. Comparison of running time of different nodes

图 5. 不同节点的运行时间对比

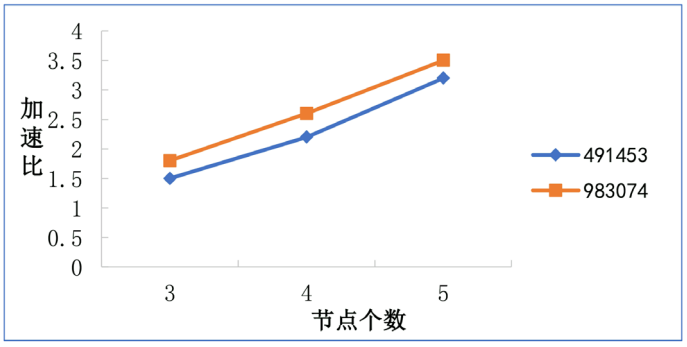


Figure 6. Acceleration ratio of different nodes

图 6. 不同节点的加速比

通过计算加速比(式(5))衡量集群运算效率，其中 sq 为加速比，t_1 代表在单机环境下所需时长，t_n 代表节点数为 n 的集群中所需的时长。比较不同节点的集群与单机环境运算效率的差异，得到不同节点的加速比对比图(图 6)，可知加速比随着节点数量和数据量的增多而增大，成正相关关系。

$$sq = t_1 / t_n \tag{5}$$

6. 结束语

本文基于 Spark 平台，结合 K-Means||算法对 GPS 出租车轨迹数据进行聚类，得到城市居民在不同时段的出行热点区域并将其直观的表示出来，分析得到各时段热点区域的热度和分布情况，与居民实际出行规律相符，且通过对比单机 K-Means 算法和在 Spark 平台上使用 K-Means||算法，结果表明后者在计算效率上优势明显。但与此同时，由于出租车仅为居民出行方式的一种，提取到的热点区域与相关特征可能不够充分，在对移动轨迹数据进行分析和挖掘的过程中，需引入其他来源的数据，使数据信息更全面、更多元，进而得出更为准确的结论和更为丰富的研究成果。

基金项目

国家自然科学基金资助项目(4156010389)。

参考文献

[1] Yue, Y., Wang, H.D., Hu, B., *et al*. (2012) Exploratory Calibration of a Spatial Interaction Model Using Taxi GPS Trajectories. *Computers*, *Environment and Urban Systems*, **36**, 140-153. https://doi.org/10.1016/j.compenvurbsys.2011.09.002

[2] Peng, C.B., Jin, X.G., Wong, K.C., Shi, M.X. and Pietro, L. (2012) Collective Human Mobility Pattern from Taxi Trips in Urban Area. *PLoS One*, **7**. https://doi.org/10.1371/journal.pone.0034487

[3] Veloso, M., Phithakkitnukoon, S. and Bento, C. (2011) Urban Mobility Study Using Taxi Traces. *International Workshop on Trajectory Data Mining and Analysis*, 23-30.

[4] 周勍, 秦昆, 陈一祥, 李志鑫. 基于数据场的出租车轨迹热点区域探测方法[J]. 地理与地理信息科学, 2016, 32(6): 51-56, 127.

[5] 张俊涛, 武芳, 张浩. 利用出租车轨迹数据挖掘城市居民出行特征[J]. 地理与地理信息科学, 2015, 31(6): 104-108.

[6] Savage, N.S., Nishimura, S., Chavez, N.E., *et al*. (2010) Frequent Trajectory Mining on GPS Data. *Proceedings of LocWeb*, ACM Press, New York, 3-7.

[7] 付鑫, 孙茂棚, 孙皓. 基于 GPS 数据的出租车通勤识别及时空特征分析[J]. 中国公路学报, 2017, 30(7): 134-143.

[8] 程静, 刘家骏, 高勇. 基于时间序列聚类方法分析北京出租车出行量的时空特征[J]. 地球信息科学学报, 2016, 18(9): 1227-1239.

[9] 牟乃夏, 张恒才, 陈洁, 张灵先, 戴洪磊. 轨迹数据挖掘城市应用研究综述[J]. 地球信息科学学报, 2015, 17(10): 1136-1142.

[10] 桂智明, 向宇, 李玉鉴. 基于出租车轨迹的并行城市热点区域发现[J]. 华中科技大学学报(自然科学版), 2012, 40(S1): 187-190.

[11] 王丽鲲. 基于社交媒体地理数据挖掘的游客时空行为分析[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.

[12] 葛小三, 付魁, 程钢, 马勇, 孙玉祥. 数据挖掘支持下的网络热点事件地理可视化研究[J]. 河南理工大学学报(自然科学版), 2016, 35(5): 655-659.

[13] 张玉峰, 曾奕棠. 基于动态数据挖掘的物流信息分析模型研究[J]. 情报科学, 2016, 34(1): 15-19, 33.

[14] 胡继华, 邓俊, 黄泽. 一种基于乘客出行轨迹的公交断面客流估算方法[J]. 计算机应用研究, 2014, 31(5): 1399-140.

[15] 毛峰. 基于多源轨迹数据挖掘的居民通勤行为与城市职住空间特征研究[D]: [博士学位论文]. 上海: 华东师范大学, 2015.

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：2161-8801，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：csa@hanspub.org