计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):64−72 [doi: 10.15888/j.cnki.csa.007308]

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

基于 Spark 的出租车轨迹处理与可视化平台①

杨卫宁, 邹维宝

(长安大学 地质工程与测绘学院, 西安 710054)
通讯作者: 杨卫宁, E-mail: yangwn@chd.edu.cn

摘　要: 大数据技术在分析与挖掘交通大数据方面扮演着越来越重要的角色. 为了快速有效地对出租车的运营模式与载客策略进行分析, 设计效益指数模型对出租车效益进行量化排序, 以高效益出租车为研究对象, 基于 Spark 大数据框架开发一个轨迹数据处理与可视化平台. 首先, 处理高效益出租车轨迹数据得到用于可视化的特征数据. 而后进行可视化分析, 包括: 统计分析高效益出租车运营特性并实现交互式图表展示, 采用蜂窝形格网与 DBSCAN 算法对不同时段高效益出租车载客点进行热点可视化, 实现基于缓冲区的交互式轨迹查询并提取出轨迹相关因子. 最后, 利用成都市出租车 GPS 轨迹数据验证了所提平台的有效性及可靠性.

关键词: 出租车轨迹; 大数据; Spark; 可视化; 城市交通

引用格式: 杨卫宁,邹维宝.基于 Spark 的出租车轨迹处理与可视化平台.计算机系统应用,2020,29(3):64–72. http://www.c-s-a.org.cn/1003-3254/7308.html

Processing and Visualization Platform of Taxi Trajectory Based on Spark

YANG Wei-Ning, ZOU Wei-Bao

(School of Geology Engineering and Geomatics, Chang'an University, Xi'an 710054, China)

Abstract: Big data technology plays an increasingly important role in analyzing and mining traffic big data. In order to quickly and effectively analyze the operating mode and passenger carrying strategy of taxis, this study designed the effectiveness index model to quantificate and sort the taxis' effectiveness. Taking high-effective taxis as the research object, a data processing and visualization platform is developed based on Spark big data framework. Firstly, high-effective taxis trajectory data are processed to obtain characteristic data for visualization. Then visual analysis is carried out, including high-effective taxis operation characteristics obtained from statistical analysis and interactive chart display, using hexagon grid and DBSCAN algorithm to visualize the hotspot of high-effective taxis carrying passenger points in different time periods, implementing interactive trajectory query based on buffer, and extracting the trajectory-related factor. Finally, the validity and reliability of this platform are verified by GPS trajectory data of Chengdu taxi.

Key words: taxi trajectory; big data; Spark; visualization; urban traffic

由于城市化进程加剧以及汽车数量增加, 城市交通问题日益严重[1], 通过分析各种空间数据解决交通问题是当前研究的热点. 出租车提供广泛且灵活的交通运输服务, 是城市交通的重要组成部分. 出租车轨迹数据记录了城市道路与居民的流动信息, 对出租车轨迹数据的挖掘分析有助于城市智慧交通[2,3]的建设, 有利于制定合理的城市交通政策、合理配置城市公共交通、缓解城市交通拥堵.

① 基金项目: 长安大学研究生科研创新实践项目 (300103002086)
Foundation item: Graduate Innovation Program of Chang'an University (300103002086)
收稿时间: 2019-08-07; 修改时间: 2019-09-05; 采用时间: 2019-09-10; csa 在线出版时间: 2020-02-28

随着经济进步与空间信息技术的发展, 出租车轨迹数据的规模呈指数级增长. 为了存储和分析生成的大量数据, 需要一种新的架构来处理出租车轨迹数据. 大数据技术的发展, 为快速、有效地处理大规模空间数据提供了可能. Spark 是 Hadoop 生态系统中新兴的杰出分布式计算框架, 具有高容错性和高可扩展性. 利用 Spark 框架的并行存储、并行计算与内存计算的优势, 可以精确有效地分析和研究城市交通问题, 实现大数据驱动的智慧交通.

高效益的驾驶员拥有丰富的驾驶经验与运营策略[4], 获取高效益的出租车轨迹数据对研究更有意义. 本文提出一种效益指数模型用于对出租车效益进行量化排序, 提取高效益出租车作为研究对象. 在此基础上设计了一个基于 Spark 的出租车轨迹处理与可视化平台, 可以快捷有效地对高效益出租车的运营模式与载客策略进行可视分析. 该平台开发对出租车轨迹数据处理和可视化具有以下贡献:

(1) 将 Spark 与 GeoTools (Java GIS 工具包) 相结合实现对出租车轨迹数据的快速处理与空间分析计算;

(2) 设计了基于蜂窝形格网与 DBSCAN 算法的出租车载客热点可视化方法, 通过时间约束, 多视角展示载客热点的变化趋势;

(3) 提出基于缓冲区的交互式轨迹查询算法, 该算法通过时间与空间约束, 将符合查询条件的轨迹数据信息可视化.

1 相关工作

1.1 出租车轨迹研究

基于位置的服务 (Location Based Services, LBS) 的快速发展, 通过出租车轨迹数据能够很好的反映出出租车的运营规律、城市交通状况以及居民出行特征. 不同领域的研究人员对出租车轨迹进行了各种研究.

在出租车运营分析方面, Weng JC 等[5]提出基于浮动车数据的出租车运营分析模型 (包括载客里程、里程利用率、驾驶员工作强度等参数). Liu L 等[4]通过分析出租车 GPS 数据了解出租车的运营模式, 对驾驶员进行分类, 揭示收益高的出租车的运营时空特征. Zhang DQ 等[6]通过挖掘出租车 GPS 轨迹, 从寻客策略、载客运营策略以及服务区域偏好三个角度研究出租车服务策略.

在出租车热点分析方面, Liu DY 等[7]开发了 SmartAdP 可视化分析系统, 利用出租车轨迹数据用于确定广告牌放置的热点区域. Chang HW 等[8]预测与时间、天气和出租车位置相关的载客需求分布, 通过 K-means、层次聚类和 DBSCAN 进行热点分析, 改善出租车运营管理. B-Planner 系统[9]使用出租车轨迹数据提取乘客上下车热点用于杭州市夜间公交路线规划.

在可视分析方面, Wang ZC 等[10]设计了一个基于 GPS 轨迹的城市交通拥堵的交互式可视化系统, 用于探索和分析城市的交通状况. 牛丹丹等[11]通过处理出租车轨迹数据, 从时间、空间维度对乘客出行特征进行可视分析. Huang XK 等[12]提出的可视化方法 TrajGraph, 通过图结构存储和可视化出租车轨迹记录的交通信息, 研究城市的交通模式.

1.2 大数据技术应用研究

大数据技术的应用, 可以快速、有效地处理大规模空间数据. 近些年出现了许多处理空间数据的分布式计算框架, 如基于 Hadoop 扩展的 SpatialHadoop[13] 与 Hadoop-GIS[14], 基于 Spark 扩展的 GeoSpark[15]、LocationSpark[16]与 Simba[17]等. 但 Hadoop MapReduce 计算模型会将中间结果输出到磁盘上, 产生大量 I/O 操作, 难以实现大规模空间数据处理. Spark 框架的性能要优于 Hadoop 框架, 通过使用 RDD, 其基于内存的并行计算架构性能更优.

在城市交通领域, 谭亮等[18]基于 Spark Streaming 和 Kafka 构建了一个实时交通数据处理平台, 处理双基基站采集的数据, 用于解决城市交通问题. 段宗涛等[19]通过 Spark 框架挖掘出租车乘客出行特征. Mao B 等[20]基于 Spark 处理、挖掘时空数据, 提出了一种基于八叉树的时空数据三维体绘制可视化框架, 对纽约市 2009–2015 年的出租车轨迹数据进行了可视化.

针对上述研究成果, 本文结合 Spark 框架的优越计算性能设计开发出租车轨迹处理与可视化平台.

2 框架与模型研究

2.1 Spark 分布式计算框架

Spark 是一个类 Hadoop 的开源分布式计算框架, 扩展了广泛使用的 MapReduce 计算模型, 用于构建大型的、低延迟的数据分析应用程序. 其主要特点是能够在内存中进行读写计算, 提升计算性能.

弹性分布式数据集 (Resilient Distributed Dataset, RDD) 是 Spark 中的基本数据抽象, 代表一个只读、可

分区、可并行计算的数据集合. RDD 可以全部或部分地缓存在内存中, 在多次计算中重用; 通过实时分发任务到所有节点, 可以保证计算的并行性. RDD 支持两种类型的操作算子: 转化操作与行动操作. 转化操作会由一个 RDD 计算生成一个新的 RDD, 行动操作会对 RDD 计算出一个结果并将结果输出 Spark 系统.

Spark SQL 是 Spark 用来处理结构化数据的一个模块, 在数据存储上采用列存的方式优化数据存储[21], 可以更便捷地处理出租车轨迹数据. Spark SQL 的核心编程抽象为 DataFrame, 一种以 RDD 为基础的分布式数据集, 记录有数据的结构信息. 同时提供 SQL 语句对数据进行操作与管理, 以 DataFrame 形式返回结果.

2.2 效益指数模型

当前城市中频繁出现的出租车拒载、空载等现象, 导致出租车运营成本增加, 城市公共交通运行效率低下, 乘客出行需求得不到满足. 针对这一问题, 如何筛选出高效益的出租车满足本文研究的数据需求, 是当前所要解决的问题.

Qin GY 等[22]发现缩短搜索时间并提高行驶速度有利于增加收入, 孙飞等[23]发现出租车单次里程长对应着单程收入高, 但若要效益高同时还要考虑寻客时间内的开销. 本文将单次载客轨迹与相邻前一段寻客轨迹相结合作为一次有效行程, 建立了出租车效益指数模型. 效益指数 F 是关于出租车的单次行程收入 I、单次里程利用率 K^1 与单次寻客时长 T^0 (min) 的函数, 有利于对出租车效益进行量化排序:

$$F = f\left(I, K^1, T^0\right) \tag{1}$$

出租车的单次行程收入 I 根据某城市的出租车计价标准确定:

$$I = V^S + \left(D^1 - D^S\right) \times V, D^1 \geq D^S \tag{2}$$

其中, V^S 表示出租车起步价格; V 表示出租车超里程单价 (RMB/km); D^S 表示出租车的起步里程 (km); D^1 表示出租车载客里程 (km).

里程利用率 K^1 是指载客里程与现实里程之比:

$$K^1 = \frac{D^1}{D^1 + D^0} \tag{3}$$

其中, D^0 表示出租车寻客里程 (km).

一般情况下, 寻客时间越长, 认为此次载客的效率越低, 本文采用寻客时长的倒数即 T^0 对效益计算加以时间控制. 可以得到效益指数的计算公式:

$$\begin{aligned} F &= I \times K^1 \times \frac{1}{T^0} \\ &= \frac{D^1 \times \left[V^S + \left(D^1 - D^S\right) \times V\right]}{T^0 \times (D^1 + D^0)}, \ D^1 \geq D^S \end{aligned} \tag{4}$$

计算某辆出租车一天的效益指数如下所示:

$$\begin{cases} \sum_{i=1}^{n} F_i = \sum_{i=1}^{n} \frac{D_i^1 \times \left[V^S + \left(D_i^1 - D^S\right) \times V\right]}{T_i^0 \times \left(D_i^1 + D_i^0\right)} \\ D_i^1 \geq D^S, \ i = 1, 2, \cdots, n \end{cases} \tag{5}$$

其中, n 表示当天该出租车的载客次数.

3 平台设计与实现

3.1 平台架构

由于浏览器软件具有高扩展、易维护、不需要安装特定软件等优势, 本文采用 B/S 软件技术架构进行平台构建, 开发了一个集大数据处理、可视化和交互于一体的“出租车轨迹处理与可视化平台”, 平台架构如图 1 所示.

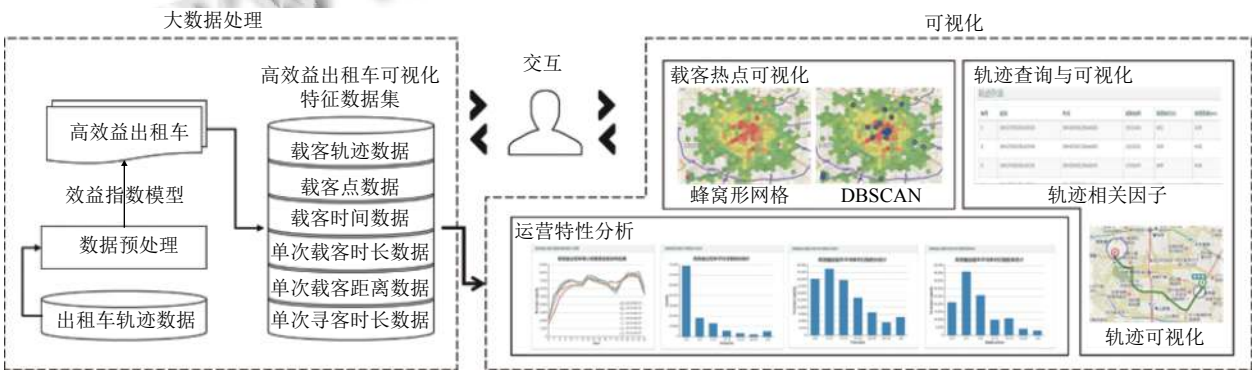


图 1 平台概览图

在大数据处理阶段, 平台基于 Spark 框架处理原始数据, 经过数据预处理及效益指数模型计算后提取出高效益出租车轨迹数据, 计算高效益出租车特征数据作为后续可视化的数据源.

可视化阶段由 3 部分组成: (1) 运营特性分析将高效益出租车特征数据进行统计计算以图表形式进行可视化; (2) 载客热点可视化结合地图数据与载客点数据, 允许用户使用蜂窝形格网与 DBSCAN 算法对不同时段高效益出租车载客点进行热点可视化; (3) 对于轨迹查询与可视化, 提取高效益出租车单条载客轨迹的轨迹相关因子并在地图上进行轨迹可视化.

3.2 基于 Spark 的轨迹大数据处理

本文利用 Spark RDD 和 Spark SQL 处理出租车轨迹大数据, 处理流程如图 2 所示, 包括数据读取与封装、数据预处理、高效益出租车提取以及可视化特征数据集计算.

数据读取与封装, 将某一天的出租车轨迹数据以字符串形式读取至 RDD 中, 构建 Taxi 类进行封装, 扩展 RDD 为 TaxiRDD, 将初始 RDD 转化为 TaxiRDD.

数据预处理, 将 TaxiRDD 导入建立的 Spark SQL 数据表中, Spark SQL 会根据 TaxiRDD 自动分配字段名称及数据类型, 使用 SQL 语句对数据进行预处理, 包括剔除异常数据、数据去重等操作.

高效益出租车提取, 将预处理完成的 DataFrame 导出为 TaxiRDD, 运用 Spark 提供的转化算子计算出租车效益指数并排序, 提取高效益出租车并输出高效益出租车轨迹数据 TaxiRDD<List<Taxi>>. 计算过程如算法 1 所示.

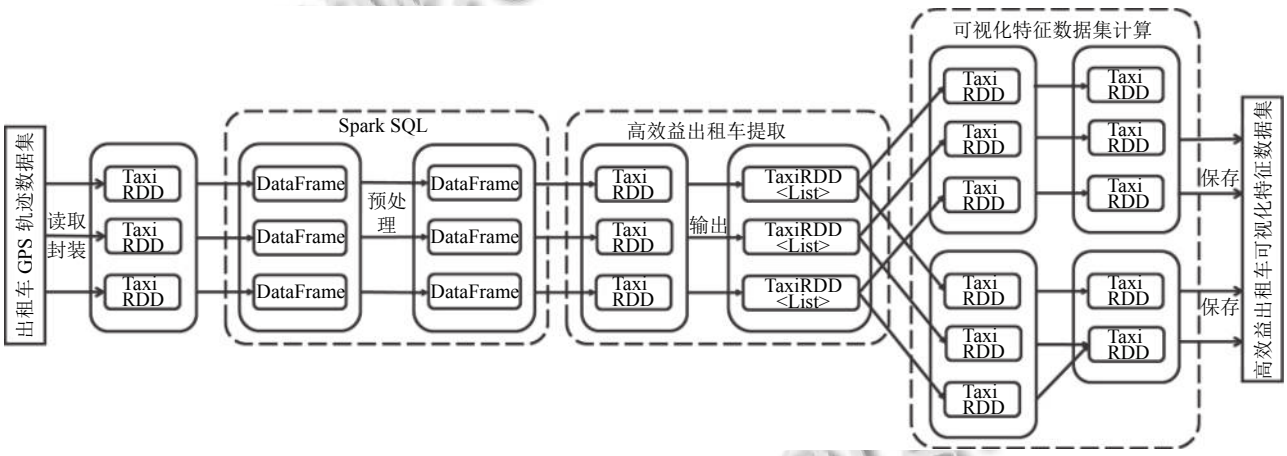


图 2 Spark 处理流程图

算法 1. 高效益出租车提取

输入: 预处理 TaxiRDD.
输出: 高效益出租车轨迹 TaxiRDD<List<Taxi>>.

1. 使用 mapToPair() 转化为键值对 TaxiRDD<出租车编号, Taxi>;
2. 使用 groupByKey() 对出租车进行分组, 再根据载客状态对轨迹进行分组;
3. 计算一次载客的效益指数;
4. 统计每辆出租车的效益指数;
5. 使用 mapToPair() 互换键值得到 TaxiRDD<效益指数, 出租车编号>, 并使用 sortByKey() 排序;
6. 使用 take() 提取高效益出租车;
7. 输出高效益出租车轨迹 TaxiRDD<List<Taxi>>.

可视化特征数据集计算, 对高效益出租车轨迹数据 TaxiRDD<List<Taxi>>中轨迹点计算得到高效益出租车可视化特征数据集. 数据集保存至本地文件夹作为后续可视分析的数据源.

3.3 运营特性分析功能

对高效益出租车的运营特性进行分析, 可以为广大出租车驾驶员的运营策略提供帮助. 运营特性主要包括高效益出租车每小时载客量分布、单次寻客时长分布、单次载客时长分布、单次载客距离分布等特性. 平台提供日期范围选择, 允许查询任意日期范围的高效益出租车运营特性. 利用 Spark 读取选中日期的可视化特征数据集, 进行统计计算, 计算结果以 JSON 格式传输至前端, 使用 ECharts[24]在浏览器界面实现交互式图表展示.

3.4 载客热点可视化功能

本功能提供日期范围选择与时间范围选择, 支持蜂窝形格网与 DBSCAN 空间聚类算法计算高效益出

租车载客热点.

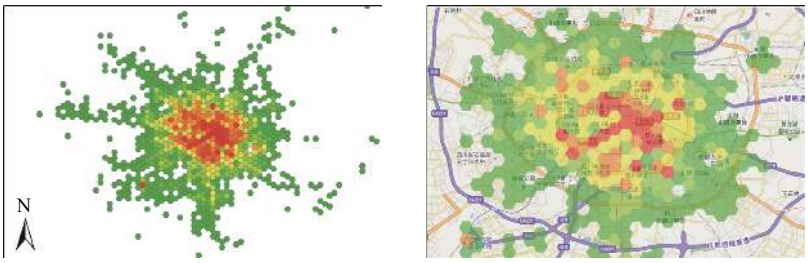
格网结构有利于分析大型空间数据集, 而蜂窝结构是覆盖二维平面的最佳拓扑结构, 六边形也是边数最多的无缝多边形. 本文采用蜂窝结构的格网可视化载客点分布状况, 实现过程如下:

(1) 利用 ArcGIS 构建蜂窝结构图层;

(2) 利用 GeoTools 将蜂窝单元读取为若干 Polygon 几何对象, 将载客点读取为若干 Point 几何对象;

(3) 使用 Spark 对 Polygon 与 Point 进行“Contains”空间拓扑运算, 计算结果生成 shp 文件 (图 3(a));

(4) 将 shp 文件使用 ArcGIS Server 进行发布, 即可在浏览器上进行可视化 (图 3(b)).



(a) “Contains” 运算结果 (b) 可视化结果

图 3 蜂窝形格网载客热点可视化

但基于网格的载客热点计算在一定程度上降低了热点计算的准确性. 故在此基础上实现了 DBSCAN 算法计算载客热点. DBSCAN 是基于密度的空间聚类算法, 能够将具有高密度的区域划分为簇, 并可在带有噪声的空间数据中发现任意形状的聚类. 设定合适的 Eps 邻域和最小包含点数 MinPts 进行聚类是 DBSCAN 算法的核心. 根据文献[25]对出租车热点区域范围的定义, 将 Eps 设定为 50 m. 在 Eps 参数确定的情况下, MinPts 取值过小将会产生过多类簇, 反之将会忽略大量非噪声对象. 结合本文研究数据, 经过多次试验, 发现 MinPts 设置为 20 最为合适.

本文在 Spark 中实现了基于 KD-Tree 最邻近搜索的 DBSCAN 空间聚类算法, 用于改进由于数据量较大造成聚类时间长的问题. 设定参数 Eps: 50 m, MinPts: 20, 计算得到聚类结果后, 将聚类结果中的每个类中心作为核心聚类点, 进行逆地址解析.

3.5 轨迹查询与可视化功能

城市交通中, 不同区域间通常存在多条可达路径, 各路径蕴含丰富的信息[26], 可用于城市道路交通分析. 本功能提供某一天的轨迹查询请求, 同时设置空间约束, 提出了基于缓冲区的交互式轨迹查询算法. 计算过程如算法 2 所示.

算法 2. 基于缓冲区的交互式轨迹查询算法

输入: 日期 *data*, 起点坐标 *oPoint*, 起点缓冲区半径 *oR*, 终点坐标 *dPoint*, 终点缓冲区半径 *dR*.

输出: 结果轨迹集合 T_D, 轨迹相关因子.

1. 根据 *data* 读取高效益出租车载客轨迹数据 T 至 Spark;
2. 根据 *oPoint*, *oR* 构建起点缓冲区 *oBuffer*;
3. 对 T 的轨迹起点进行 *oBuffer* 缓冲区运算;
4. 筛选出轨迹集合 T_O;
5. 根据 *dPoint*, *dR* 构建终点缓冲区 *dBuffer*;
6. 对 T_O 的轨迹终点进行 *dBuffer* 缓冲区运算;
7. 筛选出轨迹集合 T_D 作为结果轨迹集合;
8. 提取出 T_D 中轨迹的轨迹相关因子 (起始时间、运营时长、距离、*O*/*D* 点位置).

采用多视图协同交互的方法, 当鼠标悬停在地图上某一区域, 单击并拖动鼠标在地图上绘制 *O*/*D* 点缓冲区 (图 4), 拖动过程中控制面板文本框会显示当前绘制的 *O*/*D* 点坐标与缓冲区距离. 将控制面板中文本框内容作为输入数据. 查询结果使用天地图 JavaScript API 进行轨迹可视化.

4 实验结果与分析

4.1 数据描述

实验用到的数据集为成都市 2014 年 8 月 17 日至 23 日共 13 000 辆出租车的运营轨迹, 每天的数据记录时间为 6:00–24:00, 数据总量为 18.3 GB. 数据集以天为单位并采用 txt 格式进行存储, 每个轨迹点包含以下属性: 出租车编号, 纬度, 经度, 时间, 载客状态. 载客状态是出租车是否载有乘客的标签 (1 表示载客, 0 表示空驶).

经过效益指数模型计算, 选取 Top 20% 的出租车作为高效益出租车共计 807 267 条轨迹, 提取高效益出

租车可视化特征数据集, 并据此分析高效益出租车的运营规律.

4.2 高效益出租车运营特性分析

对 7 天的高效益出租车进行运营特性分析. 如图 5 所示, 统计各小时内的载客量, 分析高效益出租车在一天内不同时段载客量的变化特征. 在 6:00–9:00 时段载客量急剧增加, 到 10 时达到早高峰, 在 13 时达到午间需求高峰. 从 13:00–16:00 呈现缓慢下降趋势, 在 19:00–22:00 载客量又开始迅速增加, 并在 21 点达到晚高峰. 另外工作日与休息日的载客量分布略有不同. 在工作日, 13:00–18:00 载客量呈下降趋势, 并在 18 时载客量最少; 而在休息日, 13:00–18:00 载客量分布较均等.

图 4 查询缓冲区绘制

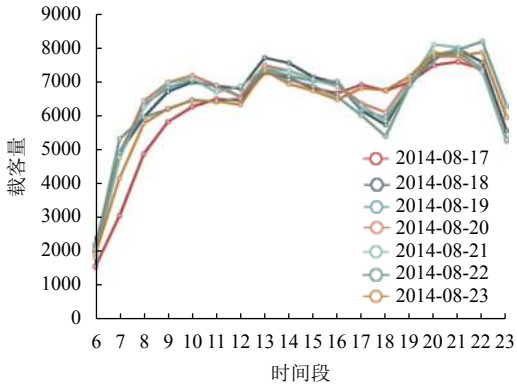


图 5 高效益出租车每小时载客量分布图 (8 月 18–22 日为工作日, 8 月 17 日与 23 日为休息日)

将 7 天的数据作均值计算, 得到高效益出租车的寻客时长、载客时长以及载客距离的分析结果. 图 6 展示了高效益出租车单次寻客时长分布结果, 发现高效益出租车寻客时长相对较短, 60% 的驾驶员会在 2 min 内寻找到乘客, 说明高效益出租车驾驶员对客源分布与道路状况非常熟悉. 图 7 展示了高效益出租车的单次载客时长分布结果, 图 8 为单次载客距离分布结果. 高效益出租车的主要服务时长在 20 min 内, 其中 15 min 以内的载客次数占总数的 66.2%, 为高效益出租车主要服务时域. 76.1% 的高效益出租车的单次载客距离多集中在 8 km 以内, 为高效益出租车主要服务半径.

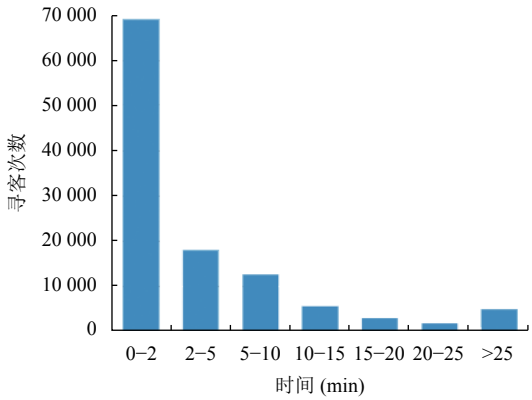


图 6 高效益出租车单次寻客时长分布图

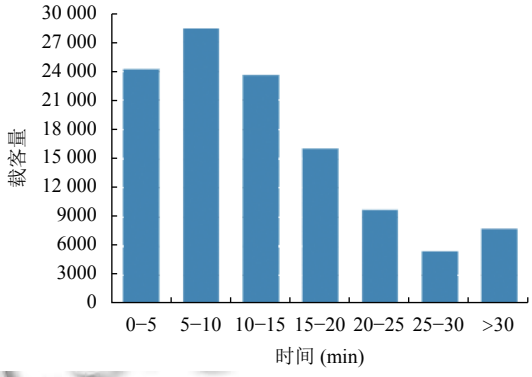


图 7 高效益出租车单次载客时长分布图

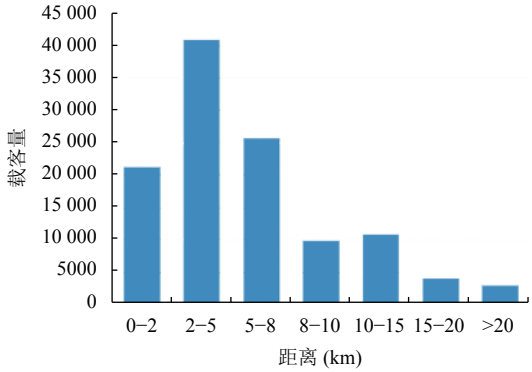


图 8 高效益出租车单次载客距离分布图

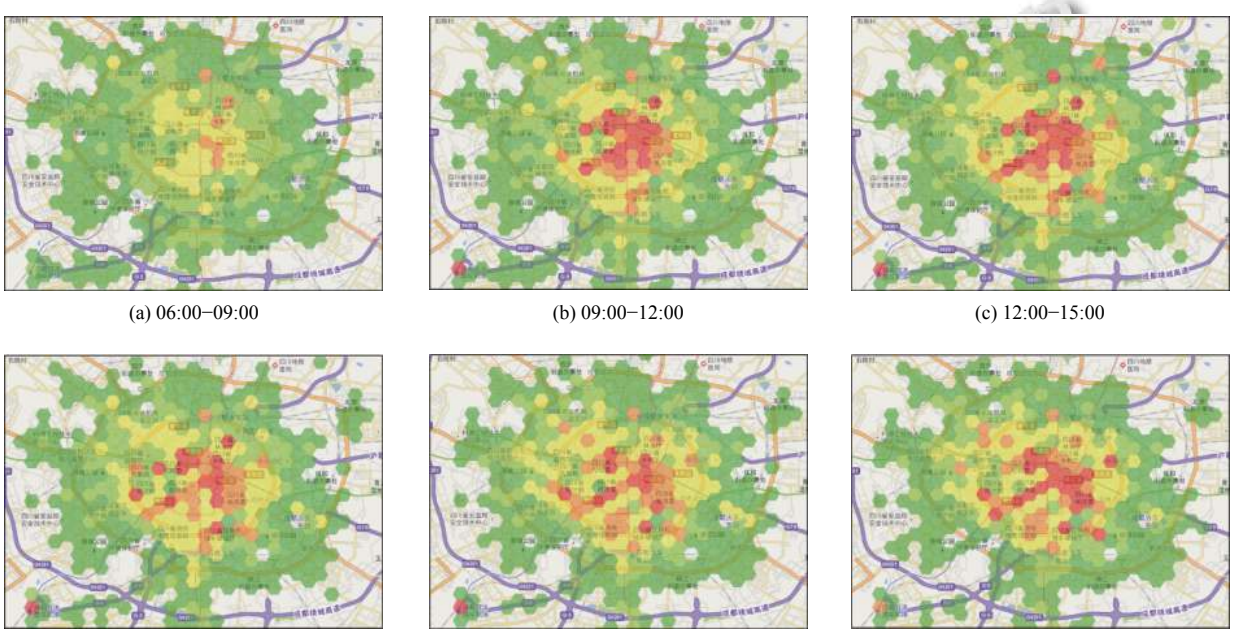
4.3 高效益出租车载客热点可视化分析

蜂窝形格网可视化可以宏观地展现城市不同时间

的载客热点变化趋势与空间分布趋势, 便于观察高效益出租车载客分布的动态过程. DBSCAN 空间聚类算法可视化可以发现高效益出租车的核心聚类点, 更细致地得到高效益出租车载客中心位置.

图 9 展示了 7 天 6 个时间段的蜂窝格网载客热点分布, 由绿到红代表载客密度的不断增大. 可以看出, 高效益出租车的载客分布大致在二环路内, 随着时间的推进向南三环路扩展. 载客热点主要分布在市中心、商场、医院、旅游景点及重要交通枢纽等地块.

图 10 展示了 21:00–24:00 时段的 DBSCAN 聚类载客热点分布, 五角星代表计算得出的核心聚类点. 点击五角星, 弹窗会显示当前核心聚类点的地址信息. 可以发现, 机场 (图 10R_a)、旅游景点 (图 10R_b)、春熙路商业区 (图 10R_d)、火车站等都属于客流集中区域. 此外, 载客热点还包括大学周边 (图 10R_c)、休闲娱乐场所 (图 10R_e) 等区域.



(a) 06:00−09:00 (b) 09:00−12:00 (c) 12:00−15:00

(d) 15:00−18:00 (e) 18:00−21:00 (f) 21:00−24:00

图 9 各时段蜂窝形格网载客热点可视化

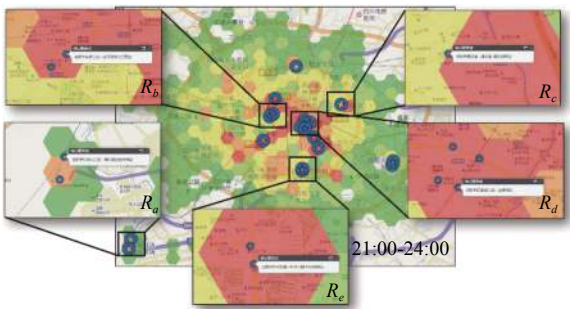
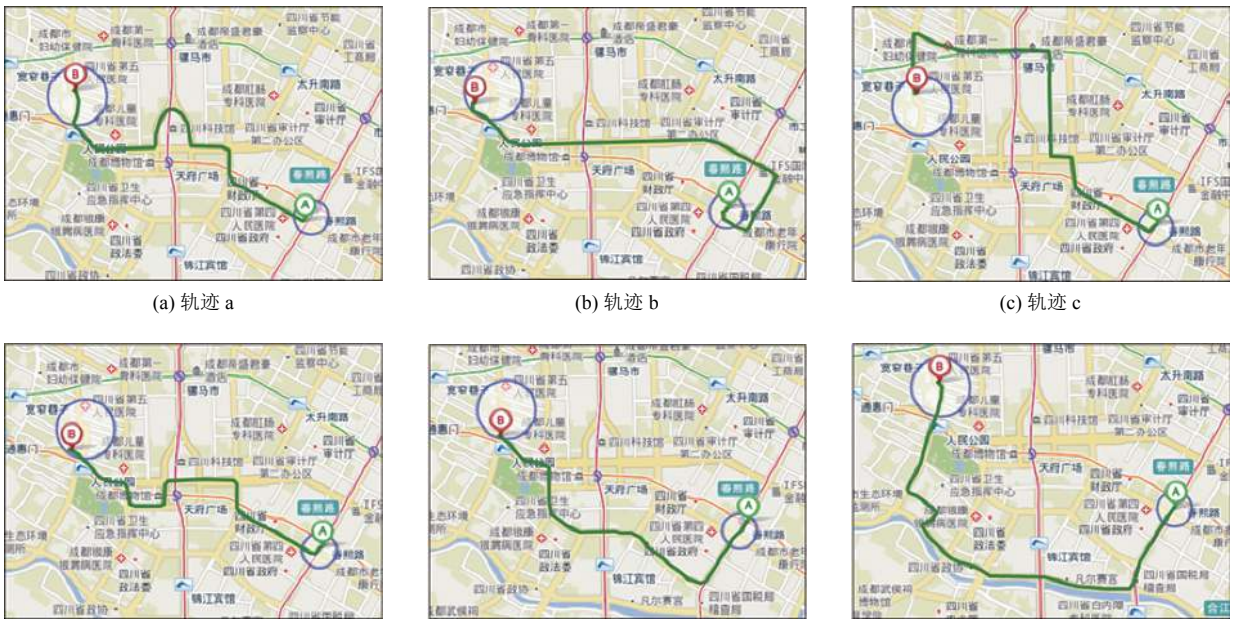


图 10 DBSCAN 聚类结果可视化

4.4 高效益出租车轨迹查询与可视化

通过在地图上绘制起点缓冲区 OB 与终点缓冲区 DB, 可视化起点缓冲区与终点缓冲区中的多条轨迹, 并提取轨迹相关因子. 选取 8 月 17 日与 23 日的数据, 将 OB 设置在春熙路地铁站附近, 缓冲区半径为 155 m, 将 DB 设置在宽窄巷子长顺上街附近, 缓冲区半径为 260 m. 共查询出 31 条轨迹, 图 11 展示了部分轨迹可视化结果. 其中, 轨迹 a、b、d、e 为高效益出租车频繁选择路径, 轨迹 c、f 只出现过一次, 为特殊路径.

轨迹可视化结合轨迹相关因子, 可以了解当前区域的道路交通状况与分析驾驶员的路径选择行为. 表 1 展示了图 11 中轨迹 a 在不同时段的轨迹相关因子, 发现 a2 的运营时长最长, a3 的运营时长最短, 说明 a 路径在午间车流量较大, 较为拥堵, 晚间车流量较小, 驾驶速度较快. 图 11 中轨迹 d、e、f 为相同时段选择不同路径的轨迹, 表 2 罗列了其轨迹相关因子, 发现轨迹 d 行驶距离最短但运营时长最长, 轨迹 f 行驶距离最长但运营时长最短. 可以发现当前时段 d 路径较为拥堵, 想要较短时间到达目的地需要选择绕行避开拥堵区域.



(a) 轨迹 a (b) 轨迹 b (c) 轨迹 c

(d) 轨迹 d (e) 轨迹 e (f) 轨迹 f

图 11 轨迹可视化 (A 标注为起点, B 标注为终点)

表 1 针对图 11(a) 中轨迹 a 不同时段相同路径的轨迹相关因子表

轨迹	起点 (lng, lat)	终点 (lng, lat)	起始时间	运营时长 (s)	距离 (km)
a1	(104.075 27, 30.655 18)	(104.053 91, 30.665 51)	17:06:47	702	3.35
a2	(104.076 49, 30.655 94)	(104.053 91, 30.665 40)	14:24:11	853	3.57
a3	(104.075 94, 30.655 06)	(104.053 98, 30.665 42)	19:49:47	686	3.46

表 2 相同时段不同路径的轨迹相关因子表

轨迹	起点 (lng, lat)	终点 (lng, lat)	起始时间	运营时长 (s)	距离 (km)
d	(104.076 34, 30.655 58)	(104.052 77, 30.663 35)	17:06:47	837	3.12
e	(104.076 87, 30.656 04)	(104.053 61, 30.662 97)	17:01:33	703	3.26
f	(104.076 18, 30.655 36)	(104.054 02, 30.665 43)	17:05:49	649	4.48

5 结束语

本文实现了一个基于 Spark 的出租车轨迹大数据处理与可视化平台, 设计效益指数模型提取高效益出租车用于可视化分析. 运营特性分析, 对研究高效益出租车运营模式、提升出租车效益具有重要意义. 对载客热点进行分析, 有利于合理配置城市公共交通、提高载客效率. 轨迹查询与可视化, 可用于城市道路交通分析、研究轨迹相关因子对路径选择行为的影响. 以成都市出租车轨迹数据作为研究实例, 验证了平台的有效性.

在未来研究中, 将继续完善平台功能, 如添加三维可视化、提供寻客推荐功能、实现实时出租车数据分析服务. 同时希望平台可以应用于不同地区, 比较不同地区出租车运营模式与载客策略的异同.

参考文献

1 Makino H, Tamada K, Sakai K, *et al*. Solutions for urban traffic issues by ITS technologies. IATSS Research, 2018, 42(2): 49–60. [doi: 10.1016/j.iatssr.2018.05.003]

2 Liu Y. Big data technology and its analysis of application in urban intelligent transportation system. Proceedings of 2018 International Conference on Intelligent Transportation, Big Data & Smart City. Xiamen, China. 2018. 17–19.

3 Gohar M, Muzammal M, Rahman AU. SMART TSS: Defining transportation system behavior using big data analytics in smart cities. Sustainable Cities and Society, 2018, 41: 114–119. [doi: 10.1016/j.scs.2018.05.008]

4 Liu L, Andris C, Ratti C. Uncovering cabdrivers' behavior patterns from their digital traces. Computers, Environment and Urban Systems, 2010, 34(6): 541–548. [doi: 10.1016/j.compenvurbsys.2010.07.004]

5 Weng JC, Zhai YQ, Zhao XJ, *et al*. Floating car data based taxi operation characteristics analysis in Beijing. Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering. Los Angeles, CA, USA. 2009. 508–512.

6 Zhang DQ, Sun L, Li B, *et al*. Understanding taxi service strategies from taxi GPS traces. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 123–135. [doi: 10.1109/TITS.2014.2328231]

7 Liu DY, Weng D, Li YH, *et al*. SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 1–10.

8 Chang HW, Tai YC, Hsu JYJ. Context-aware taxi demand hotspots prediction. International Journal of Business Intelligence and Data Mining, 2010, 5(1): 3–18. [doi: 10.1504/IJBIDM.2010.030296]

9 Chen C, Zhang DQ, Li N, *et al*. B-Planner: Planning bidirectional night bus routes using large-scale taxi GPS traces. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(4): 1451–1465. [doi: 10.1109/TITS.2014.2298892]

10 Wang ZC, Lu M, Yuan XR, *et al*. Visual traffic jam analysis based on trajectory data. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2159–2168. [doi: 10.1109/TVCG.2013.228]

11 牛丹丹, 段宗涛, 陈柘, 等. 城市出租车乘客出行特征可视化分析方法. 计算机工程与应用, 2019, 55(6): 237–243. [doi: 10.3778/j.issn.1002-8331.1712-0019]

12 Huang XK, Zhao Y, Ma C, *et al*. TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 160–169. [doi: 10.1109/TVCG.2015.2467771]

13 Eldawy A, Mokbel MF. Spatialhadoop: A mapreduce framework for spatial data. Proceedings of the IEEE 31st Interna-tional Conference on Data Engineering. Seoul, Republic of Korea. 2015. 1352–1363.

14 Aji A, Sun XL, Vo H, *et al*. Demonstration of Hadoop-GIS: A spatial data warehousing system over MapReduce. Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Orlando, FL, USA. 2013. 528–531.

15 Yu J, Zhang ZS, Sarwat M. Spatial data management in Apache Spark: The GeoSpark perspective and beyond. GeoInformatica, 2019, 23(1): 37–78. [doi: 10.1007/s10707-018-0330-9]

16 Tang MJ, Yu YY, Malluhi QM, *et al*. LocationSpark: A distributed in-memory data management system for big spatial data. Proceedings of the VLDB Endowment, 2016, 9(13): 1565–1568. [doi: 10.14778/3007263.3007310]

17 Xie D, Li FF, Yao B, *et al*. Simba: Efficient in-memory spatial analytics. Proceedings of the 2016 International Conference on Management of Data. San Francisco, CA, USA. 2016. 1071–1085.

18 谭亮, 周静. 基于 Spark Streaming 的实时交通数据处理平台. 计算机系统应用, 2018, 27(10): 133–139. [doi: 10.15888/j.cnki.csa.006592]

19 段宗涛, 陈志明, 陈柘, 等. 基于 Spark 平台城市出租车乘客出行特征分析. 计算机系统应用, 2017, 26(3): 37–43. [doi: 10.15888/j.cnki.csa.005617]

20 Mao B, Yu ZC, Cao J. Large scale spatial temporal data visualization based on Spark and 3D volume rendering. Proceedings of 2016 International Joint Conference on Neural Networks. Vancouver, BC, Canada. 2016. 1879–1882.

21 李格非, 马蔚吟, 李力. Spark 平台下的凸包问题研究. 计算机工程与应用, 2018, 54(22): 67–73, 112. [doi: 10.3778/j.issn.1002-8331.1708-0293]

22 Qin GY, Li TN, Yu B, *et al*. Mining factors affecting taxi drivers' incomes using GPS trajectories. Transportation Research Part C: Emerging Technologies, 2017, 79: 103–118. [doi: 10.1016/j.trc.2017.03.013]

23 孙飞, 张霞, 唐炉亮, 等. 基于 GPS 轨迹大数据的优质客源时空分布研究. 地球信息科学学报, 2015, 17(3): 329–335.

24 Li DQ, Mei HH, Shen Y, *et al*. ECharts: A declarative framework for rapid construction of web-based visualization. Visual Informatics, 2018, 2(2): 136–146. [doi: 10.1016/j.visinf.2018.04.011]

25 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. [doi: 10.15888/j.cnki.csa.006863]

26 Lu M, Lai CF, Ye TZ, *et al*. Visual analysis of multiple route choices based on general GPS trajectories. IEEE Transactions on Big Data, 2017, 3(2): 234–247. [doi: 10.1109/TBDATA.2017.2667700]