基于出租车 GPS 数据的商圈分析

朱昶胜 刘敬帅 李 硕
(兰州理工大学计算机与通信学院 甘肃 兰州 730050)

摘 要 出租车运营特性具有随机性、即走即停、覆盖范围广的特点,行驶的起止点由乘客决定,其运营规律能够很好地反映乘客出行的特点。根据出租车 GPS 的定位数据在真实地理空间的覆盖情况,可以还原居民出行的活动轨迹,挖掘潜在信息。提出采用出租车 GPS 定位数据进行商圈分析。通过对 GPS 定位数据进行网格划分、聚类,使用 R 语言建立相应的数据模型以及对模型的应用和结果分析。实验结果表明,将不同时段、不同地点的出租车特征进行统计分析做出折线图,可以识别出不同商圈类型,根据这些信息为潜在顾客的分布制定适宜的商业对策。

关键词 数据挖掘 R 语言 GPS 数据 商圈分析

中图分类号 TP391 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2017.10.021

ANALYSIS OF BUSINESS CIRCLE BASED ON TAXI GPS DATA

Zhu Changsheng Liu Jingshuai Li Shuo
(*School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, Gansu, China*)

Abstract Taxi operation characteristics are random, with flexible boarding area, covering a wide range. Because the starting and ending points are determined by the passengers, therefore, taxi's operating rules can reflect the characteristics of passengers travel well. According to the GPS data in the real geographical space coverage, we can restore the activities of residents travel trajectory, and dig out the potential information. So we propose the use of taxi GPS location data for business district analysis. Through meshing and clustering with GPS location data, corresponding mathematical models were established with R language and the application and results analysis of the model were discussed. The experimental results show that through the statistical analysis to characteristics of taxi, those at various times and in different locations, the line chart can be built. From this chart, the types of business district can be identified. According to this information, the suitable business strategy can be made for the potential customer's distribution.

Keywords Data mining R language GPS Data Business circle analysis

0 引 言

近年来,随着城市化进程和机动车保有量的迅猛增长,道路交通在为人们出行带来便捷的同时也带来了交通事故、交通拥挤、交通污染等诸多负面的影响。出租车作为交通系统的组成部分之一,装有 GPS 的出租车系统能够提供路网交通状态的时变数据和及时准确的运营数据。通过对这些数据的应用分析不但有助于合理地缓解城市交通拥堵,从根本上提高出租车行业的整体服务水平,而且能了解居民日常出行行为及居民个性化的服务需求,同时为交通规划及城市管理提供决策支持。因此对于出租车 GPS 数据应用的深入研究有重要的现实意义[1]。

收稿日期:2016-10-20。甘肃省自然科学基金项目(148RJZA019);甘肃省高校科研项目(2015B-031)。朱昶胜,教授,主研领域:大数据,云计算。刘敬帅,硕士。李硕,硕士。

对于出租车 GPS 数据的应用,国内外学者主要从四个方面进行研究: 交通状态估计研究、交通行为研究、出行 OD 预测研究及出租车运营管理研究。虽然产生了一定的研究成果,但同时也存在一定的不足。例如在数据分析方面仅限于运营信息的提取研究,没有对这些信息背后的潜在信息进行挖掘等。因此,随着信息技术的发展,有必要对 GPS 数据进行深入的挖掘。

商圈是现代市场中企业市场活动的空间,最初是站在商品和服务提供者的产地角度提出来的,后来逐渐扩展到商圈同时也是商品和服务享用者的区域。商圈划分目的之一是研究潜在的顾客的分布以制定适宜的商业对策[2]。

R 语言是一个用于统计计算和统计制图的优秀工具,具有高效的数据建模、统计分析及可视化能力[3]。

本文基于出租车 GPS 数据的时变和及时准确的特性,利用出租车随机性、覆盖范围广的特点,使用 R 语言来对 GPS 数据进行挖掘来进行商圈区域的划分[4]。

1　基于 GPS 数据的商圈分析框架

出租车作为城市交通系统的组成部分,其运营特性具有随机性、即走即停、覆盖范围广的特点,行驶的起止点由乘客决定,其运营规律能够很好地反映乘客出行的特点。出租车 GPS 数据主要由经度、纬度、速度、时间、载客状态、方向等。定位数据描绘了居民出行的活动模式,通过对定位数据的分析识别出不同类别的热点地区,即可识别出不同类别的商圈。衡量区域的特征可以从出租车流量和上下客数量的角度进行分析,所以在归纳热点区特征时可以从这两个特点进行提取[5]。

基于 GPS 数据的商圈分析如图 1 所示,主要包括以下步骤:

1) 对原始出租车 GPS 数据预处理,剔除数据,提取分析所需的期望数据。

2) 选取一周的数据及特定的区域,通过网格划分来进行分块处理,研究不同时间段(工作日、节假日) 各分块内停留次数。

3) 对数据进行数据规约和数据变换处理,建立数据分析模型,基于网格分块区域的出租车特征进行商圈聚类。

4) 对各个商圈分群进行特征分析,对不同区域提出合理化规划建议。

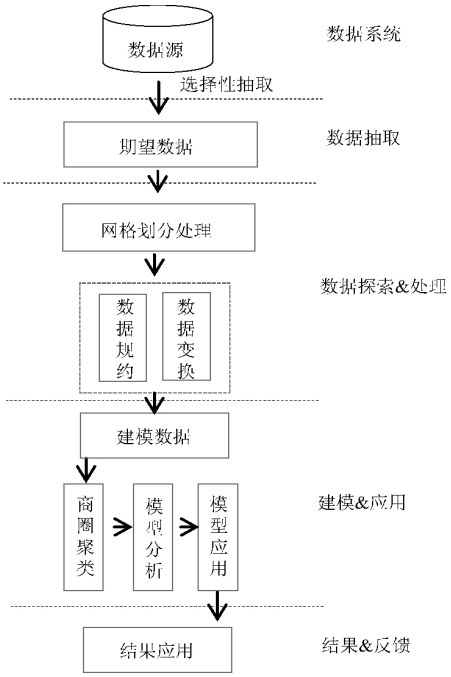


图 1　GPS 数据的商圈分析流程图

2　数据抽取

在实际应用中,GPS 采样信号的质量会由于采样频率的降低、定位误差的加大、信号的丢失等的影响,需要对其进行消除数据冗余、缺失和错误的情况。本文选取一周的出租车历史数据进行处理分析,在数据抽取阶段完成对数据的预处理。原始数据中包含经度、纬度、发送时间、接收时间、速度、方向、状态等记录信息,出租车 GPS 的数据格式如表 1 所示,本文采用的数据来自甘肃天水的出租车数据,使用的数据为 600 辆出租车一周的数据。

表 1　GPS 数据格式

ID	时间	经度	纬度	速度	方向	ACC
297 100	2014-07-29	101.857 54	36.578 22	32	122	1

原始数据的属性较多,对于商圈数据的挖掘并不需要这么多数据,因此在数据预处理阶段将目标数据给提取出来。对于商圈具有客流量大、上下客多的特点,某一天的数据无法判定某些地区是否是人流密集区。因此在这里选取一周的数据来进行分析: 周一至周五为工作时间,周六至周日是节假日时间,通过不同时间段来对数据进行挖掘分析。

3　数据探索分析

由于出租车的特性所造成的出租车运行轨迹的随

机性比较强、覆盖范围比较广，致使城市中的任何地方都有 GPS 定位点。对于商圈划分来说，首先的要求就是定位的热点地区，因此，首先应该对 GPS 定位数据进行分类划分，将小于一定阈值的定位点给清除，保留定位点丰富的地区[6-7]。

3.1 数据提取分析

GPS 数据的采集具有规律性，其定位数据信息每隔 30 秒会上传一次，当速度不为零时，状态位显示为 0，当速度为 0 时，状态位显示为 1。通过数据清洗，若出现连续的为 0 状态则将 0 的状态剔除，仅保留一个数据为 0 状态。速度为 1 时可看作是载客状态。故可以设定当状态参数从 0 变为 1 或从 1 变为 0 时可作为上下客的点。通过统计状态为 1 的数据的数量，即可得到总的上下客的数量。

3.2 基于网格的划分

基于网格的划分是将对象空间量化为一定数目的单元格，形成一个网络结构，然后依次统计每个网格内的数据量，之后进行聚类，所有的聚类都在这个网格上进行[8]。在这里采用 STING 聚类算法。STING 是一种基于网格的多分辨率聚类技术，它将空间划分为矩形单元。

对每个网格中的上下客的量进行统计，将小于阈值的数据清除，然后进行聚类分析，如图 2 所示。对网格中的数据进行按照不同的时段进行分类，统计出各个时段的数据量用于数据分析。

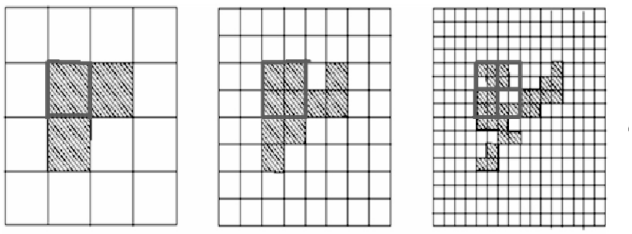


图 2 网格划分

3.3 数据统计分析

为了寻找高价值的商圈，要根据定位数据提取出相应区域的客流量特征，如上下客数量、车流量等。高价值的商圈具有客流量、车流量大的特点，但是一些区域是工作日时间乘客出行较多，一些区域是周末出行较多，而有些地方则是晚间的出行较多，所以提取的特征必须明显地区别这些区域。下面设计工作日上班时间上下客数量、夜间上下客数量、周末上下客的数量和车流量做为特征进行分析[9]。

本文中将工作日上班时间定为 8:00 – 18:00，夜间时间为 18:00 – 24:00，周末则是在相应区域的总量，在商圈区域周末的客流量与车流量将会大幅增加。车流量是指在该区域内有 GPS 定位即可表示为有车在此经过。这个时间段比较符合人们的工作生活需要，因此以这些特点进行统计。对一定区域的上下客客流特征的计算公式如下:

上班时间: $$weekday = \frac{1}{5}\sum_{j=1}^{5} weekday_j \quad (1)$$

夜间时间: $$night = \frac{1}{5}\sum_{j=1}^{5} night_j \quad (2)$$

周末时间: $$weekend = \frac{1}{5}\sum_{j=1}^{5} weekend_j \quad (3)$$

车流量: $$flow = \frac{1}{5}\sum_{j=1}^{5} flow_j \quad (4)$$

3.4 数据标准化处理

由于四个特征值的差异较大，为了消除不同属性间的不齐性，需要对数据进行离差标准化。离差标准化，是对原始数据的线性变换，使结果落到 [0,1] 区间，转换函数如下:

$$X_{ij}^{*} = \frac{X_{ij} - \bar{X}_j}{s_j} \quad i = 1,2,\cdots,n \quad j = 1,2,\cdots,p \quad (5)$$

其中 $\bar{X}_j = \frac{1}{n}\sum_{k=1}^{n} X_{kj}$，$S_j = \frac{1}{n-1}\sum_{k=1}^{n}(X_{kj} - \bar{x}_j)^2$。

每个变量的样本均值为 0，标准差为 1，而且标准化后的数据与变量的量纲无关。

4 建模与应用

4.1 距 离

设 X_{ik} 为第 i 样本的第 k 个指标，每个样本有 p 个变量，第 i 个样本与第 j 个样本之间的距离记为 d_{ij}，在聚类过程中，距离较远的归为一类，距离较近的归为一类，定义的距离满足以下四个条件:

$d_{ij} \geqslant 0$，对一切 i,j;

$d_{ij} = 0$，当且仅当第 i 个样本与第 j 个样本的个变量值相同;

$d_{ij} = d_{ij}$，对一切 i,j;

$d_{ij} = d_{ik} + d_{kj}$，对一切 i,j,k;

本文使用 Euclide 距离，公式如下:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2} \quad (6)$$

4.2 离差平方和聚类

离差平方和法基于方差分析思想，如果类分得正确，则同类样本直接的离差平方和应当较小，不同类样本之间的离差平方和应当较大。

设 d_{ij} 表示第 i 个样本与第 j 个样本的距离，G_1，G_2，

… 表示类，D_{KL} 表示 G_K 和 G_L 的距离。类 G_K 和 G_L 合并成新的类 G_M，则 G_K，G_L，G_M 的离差平方和分别是:

$$W_K = \sum_{i \in G_K} (x_{(i)} - \bar{x}_K)^{\mathrm{T}} (x_{(i)} - \bar{x}_K) \quad (7)$$

$$W_L = \sum_{i \in G_L} (x_{(i)} - \bar{x}_L)^{\mathrm{T}} (x_{(i)} - \bar{x}_L) \quad (8)$$

$$W_M = \sum_{i \in G_M} (x_{(i)} - \bar{x}_M)^{\mathrm{T}} (x_{(i)} - \bar{x}_M) \quad (9)$$

其中 $\bar{x}_K$、$\bar{x}_L$、$\bar{x}_M$ 分别是 G_K、G_L、G_M 的重心，所以 W_K、W_L、W_M 反映了各自类内部样本的分散程度，如 G_K 和 G_L 这两类相聚较近，则合并后所增加的离差平方和 $W_M - W_K - W_L$ 应较小，否则应较大。于是定义 G_K 和 G_L 之间的平方距离为:

$$D_{KL}^2 = W_M - W_K - W_L \quad (10)$$

这种系统聚类法称为离差平方和法[10]。通过该方法对经过网格划分处理后的数据进行利差和聚类，将有相似规律的区域进行分类，得到不同类别区域的谱系图，如图 3 所示。

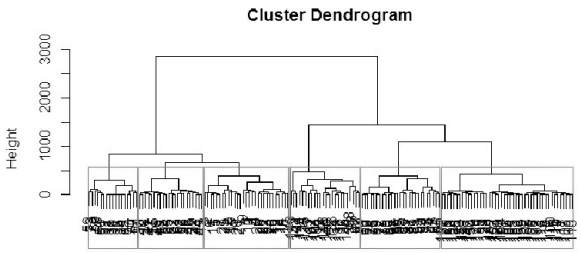
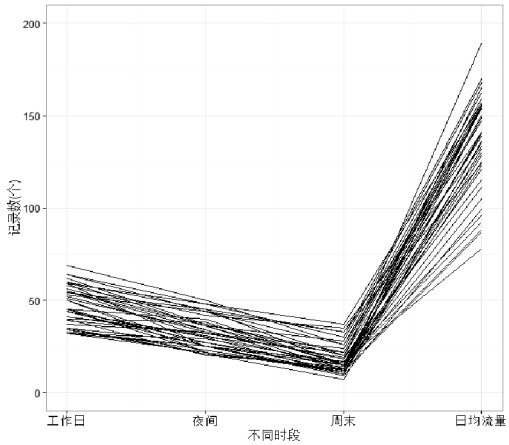


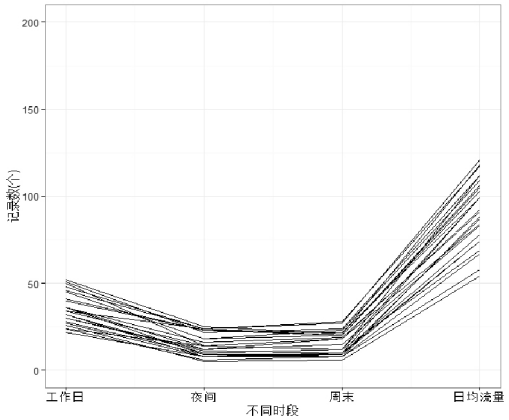
图 3　谱系聚类图

4.3　模型分析

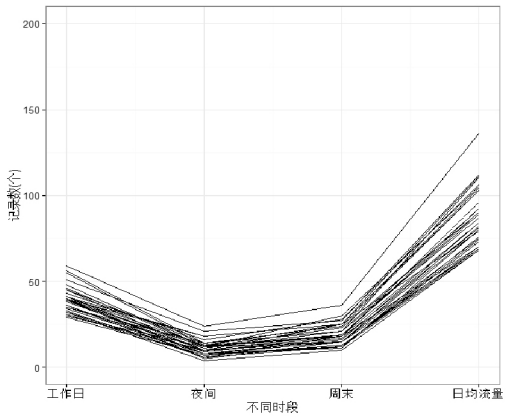
通过图 3 的聚类图可以看出，数据根据离差平方和法分为 6 类，即出租车的使用情况可以分为 6 种模式，使用 R 语言通过对聚类算法按不同类别分别画出 6 类特征的折线图，如图 4 所示。其中一条折线表示 3.2 节中的网格划分过程中的一个区域，同一个图表中的所有折线是具有相似特性的区域的集合。



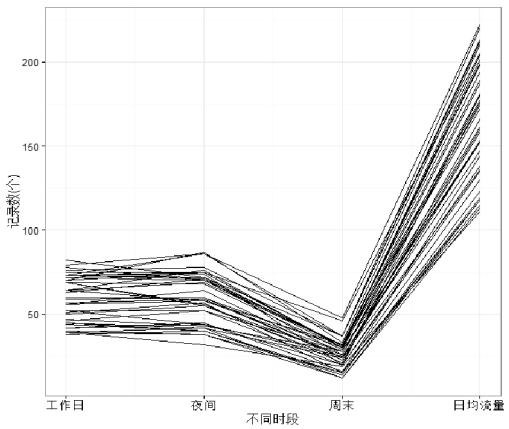
(a)



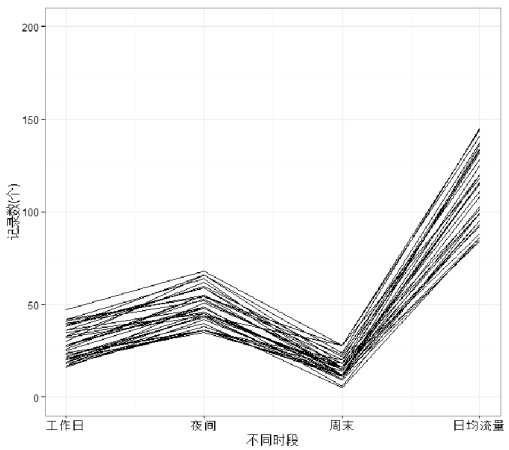
(b)



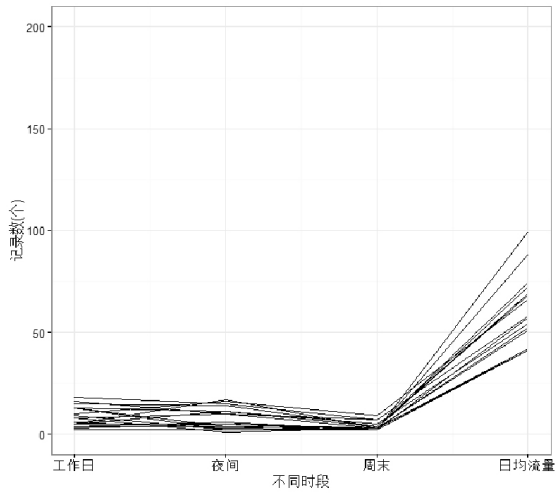
(c)



(d)



(e)



(f)

图 4 分析结果

图 4 中,(a) 为该区域工作日上班时间上下客数量多于夜间时的乘客数量,周末的乘客数量最低,该区域类似于上班的区域,工作日时间客流量较大,因此可以根据上班族的需要提供相关的服务;(b) 为该区域的周末和夜间的上下客数相似,工作日时间也比较高,该区域类似于住宅小区等区域;(c) 为夜间的上下客数量少于周末和工作日的数量,该区域夜间的上下客数量较少,白天客流量较多,类似于儿童乐园等区域;(d) 为白天和夜间的上下客数量相似,周末的数量相对也较多,该区域适合大型商场、超市等;(e) 为夜间上下客数量较多,说明该区域更偏向夜间活动,类似于夜场等地区;(f) 为三个时间段的上下车数量较少,客流量也相对较少,相比较而言该区域比较偏僻,不适合做为商业开发。

5 结 语

(1) 本文结合 R 语言和车载 GPS 定位数据的优点提出了一种 R 环境 + GPS 定位数据进行商圈分析的方法。

(2) 本实验通过对 GPS 数据进行预处理并获取样本数据,使用 R 语言对样本数据建模得到进行商圈分析的模型。

(3) 本实验采用网格划分和离差平方和聚类对出租车 GPS 数据进行建模,并通过实验验证和比较得到分析结果。

(4) 实验结果表明,对于不同时段、不同地点的数据的分析结果呈现出不同类别,在 GPS 数据中挖掘出人口空间分布和活动的特征,可以根据分析结果来进行商业圈的划分同时也可以对城市规划、出租车的调度提供合理化的建议等。

参 考 文 献

[1] 张红,王晓明,朱昶胜,等. 基于大数据的智能交通体系架构[J]. 兰州理工大学学报,2015,41(2):113-114.

[2] 孟诗琼,孟诗瑶,尹至. 基于 R 语言的汽车消费数据挖掘及可视化方法[J]. 宁波工程学院学报,2015,27(4):19-21.

[3] 杨霞,吴东伟. R 语言在大数据处理中的应用[J]. 科技资讯,2013(23):19-20.

[4] 齐林. 基于 GPS 数据的出租车交通运行特性研究及应用[D]. 哈尔滨:哈尔滨工业大学,2013.

[5] 张建发. GIS 技术在商圈分析中的应用[J]. 赤峰学院学报(自然科学版),2012,28(2):58-59.

[6] 张健钦,楸培元,杜明义. 基于时空轨迹数据的出行特征挖掘方法[J]. 交通运输系统工程与信息,2014,14(6):73-74.

[7] 郑运鹏,赵刚,刘健. 基于出租车 GPS 数据的交通热区识别方法[J]. 北京信息科技大学学报,2016,31(1):31-32.

[8] 赵慧,刘希玉,崔海清. 网格聚类算法[J]. 计算机技术与发展,2010,20(9):84-85.

[9] 张用川. 基于手机定位数据的用户出行规律分析[D]. 昆明:昆明理工大学,2013.

[10] 张良均,云伟标,王路,等. R 语言数据分析与挖掘实战[M]. 北京:机械工业出版社,2015.

(上接第 112 页)

[8] Meskini S, Nassif A B, Capretz L F. Reliability models applied to mobile applications[C]//Software Security and Reliability-Companion (SERE-C), 2013 IEEE 7th International Conference on. IEEE, 2013:155-162.

[9] 丁建立,张超,王静. 民航旅客服务信息系统报警规则提取[J]. 计算机仿真,2015,32(2):83-86.

[10] Kanoun K, Laprie J C. Software reliability trend analyses from theoretical to practical considerations[J]. IEEE Transactions on Software Engineering, 1994, 20(9):740-747.

[11] Yamada S, Osaki S. Software reliability growth modeling: Models and applications[J]. IEEE Transactions on Software Engineering, 1985, 11(12):1431.

[12] Zhang Xuemei, Hoang P. Comparisons of nonhomogeneous Poisson process software reliability models and its applications[J]. International Journal of Systems Science, 2000, 31(09):1115-1123.