·智能交通专题· 文章编号:1000-3428(2017)05-0016-07 文献标志码:A 中图分类号:TP39

基于载客数据的出租车热门区域功能发现

孙冠东[1,2],张 兵[1,2],刘禹岍[1,2],熊 赟[1,2]

(1. 复旦大学 计算机科学技术学院,上海 201203;2. 上海市数据科学重点实验室,上海 201203)

摘 要: 根据出租车行驶载客数据中提取的乘客出行模式和上下客热门区域,提出一种出租车热门区域功能发现方法。采用基于交通数据时空特性的出租车行驶数据聚类算法,实现热门区域划分。建立基于潜在 Dirichlet 分配的热门区域乘客出行特征发现模型,对具有相似乘客出行模式的出租车热门区域进行聚类。通过总结各热门区域的具体功能,发现在不同客流时间段内的区域功能与乘客出行模式间的关系。实验结果表明,该方法能够有效发现热门区域的功能特点。

关键词: 时空特性;区域功能;热门区域发现;主题模型;乘客出行模式

中文引用格式: 孙冠东,张 兵,刘禹岍,等. 基于载客数据的出租车热门区域功能发现[J]. 计算机工程,2017,43(5):16-22.

英文引用格式: Sun Guandong, Zhang Bing, Liu Yuqian, et al. Taxi Hot Area Function Discovery Based on Passenger Data[J]. Computer Engineering, 2017, 43(5):16-22.

Taxi Hot Area Function Discovery Based on Passenger Data

SUN Guandong[1,2], ZHANG Bing[1,2], LIU Yuqian[1,2], XIONG Yun[1,2]

(1. School of Computer Science, Fudan University, Shanghai 201203, China;
2. Shanghai Key Laboratory of Data Science, Shanghai 201203, China)

【Abstract】 According to the passenger movement pattern and the hot pick-up and drop-off areas extracted from taxi driving passenger data, this paper proposes a functions discovery method of taxi hot areas. Firstly, it uses taxi driving data clustering algorithm based on the temporal and spatial characteristics of traffic data to realize hot region division. Then, the passengers travel character discovery model of passengers in hot region based on Latent Dirichlet Allocation(LDA) is built to realize clustering hot taxi region with similar passenger travel mode. Finally, by summarizing the specific function of each area, it can find the relationship between area function and passenger movement patterns at different period of passenger flow. The experimental results show the method can effectively discovery the function characteristics of hot areas.

【Key words】 temporal and spatial characteristics; area function; hot area discovery; topic model; passenger travle mode

DOI: 10.3969/j.issn.1000-3428.2017.05.003

0 概述

出租车已成为城市出行的重要交通工具之一,具有方便、快捷、舒适等特点。随着智能交通系统的快速发展,出租车轨迹数据积累越来越丰富。发现出租车热门区域并了解其区域功能,对完善城市交通规划有重要的指导意义,并且能为乘客出行和司机载客提供帮助,提高资源利用率,降低空载率。

关于出租车热门区域发现问题,文献[1]通过分析出租车服务的乘车模式,为空载出租车提供位置推荐。文献[2]通过上下文感知对出租车热门区域进行预测,并设计了一个简单的应用系统。文献[3]对车辆的速度信息和轨迹信息进行基于密度的聚类,从而发现一些感兴趣的区域。文献[4]用最近邻聚类算法对乘客上车点与下车点进行聚类分析,并对不同区域的吸引力进行分析。文献[5]对乘客上车点与下车点坐标间的关系进行研究。文献[6]使用网格分解对数据进行处理,利用朴素贝叶斯预测

基金项目: 国家自然科学基金(71331005);上海市科委基金(14511107302,16511102204);NSFC-广东联合基金(第二期)超级计算科学应用研究专项;国家超级计算广州中心基金。

作者简介: 孙冠东(1992—),男,硕士研究生,主研方向为智能交通、数据挖掘;张 兵,博士研究生;刘禹岍,硕士研究生;熊 赟,教授。

收稿日期: 2016-05-31 **修回日期:** 2016-07-04 **E-mail:** gdsun15@fudan.edu.cn

不同区域的空载出租车数量。虽然上述研究能够用于发现热门区域,方便出租车的载客选择,但对城市规划而言,还需要做进一步研究。对于热门区域,需进一步解释其功能,为政府规划决策和出租车管理政策制定提供辅助支持。

对于区域功能发现这一问题,文献[7-8]提取了区域轨迹数据特征,通过手工标注部分区域的功能,建立基于 SVM 的分类模型对区域功能进行划分,同时还利用轨迹数据对乘客出行进行预测。文献[9]提出按照居民的移动性和 POIs 将北京市划分成多个功能区域。文献[10]利用地铁客流数据进行区域功能聚类。一般地,某一区域内的乘客出行模式在一定程度上可以反映该区域的功能,本文提出基于出租车数据的热门区域及其区域功能的挖掘方法。

1　出租车热门区域发现

本节通过对出租车客流分布,采用基于密度的聚类算法实现热门区域划分。选择基于密度的聚类方法是由于对于出租车热门区域发现问题,簇的数量较难预先确定并且簇可能是非球状的,因此相对于基于划分的聚类方法(如 K-means 算法),基于密度的聚类算法更加合适应用于热门区域划分[11],并且可以有效地避免噪声影响[12]。但是由于不同时段出租车的上下车人数具有较大差距,对不同时段采用统一的参数(例如 DBSCAN 算法中的 *minpts* 和 *eps*[13])进行热门区域发现并不合理,因此,不能直接使用基于密度的聚类算法。下面分析出租车客流峰段以及客流区域分布情况,在此基础上提出基于交通数据时空性的出租车数据聚类方法。

1.1　出租车客流峰段划分

出租车数据作为典型的交通数据,具有明显的时空特性。图 1 是对上海市 2015 年 4 月份出租车行驶数据在不同时间段的人数统计结果。

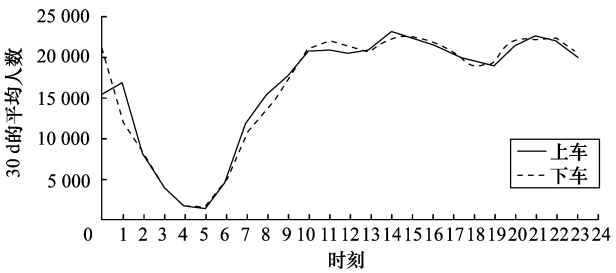

图 1　不同时刻上下车人数统计

由图 1 可以发现,在一天当中上车和下车人数在全天的趋势是大致相同的,但在某些时段两者是不一致的(例如 00:00:00—01:00:00),这说明在不同时段出租车的乘车需求程度存在差异。

根据图 1 的客流量趋势以及数量将时间段划分如表 1 所示,例如 01:00:00 开始上下车人数开始下降,直到 05:00:00,并且 05:00:00 到 06:00:00 之间出现人数仍在低点,即使 05:00:00 开始已经呈现上升趋势,因此,考虑到一般的居民生活工作情况,将(01:00:00—05:59:59)作为第 1 个时间段,其他时间段类推。

表 1　客流时间段划分

时间段	开始时间	结束时间
1	01:00:00	05:59:59
2	06:00:00	09:59:59
3	10:00:00	16:59:59
4	17:00:00	19:59:59
5	20:00:00	00:59:59

1.2　出租车分布情况

从图 1 中可以看出,因为不同时段内出租车的上下车人数有很大差距,人数最多时有 23 000 人,而人数最少时只有 1 400 人,所以采取统一的参数进行热门区域发现是不合理的。本文采用网格划分的方式将城市按经纬度进行划分(经度范围 31.100° ~ 31.325°,纬度范围 121.310° ~ 121.615°),以便发现其分布规律。网格划分的方法便于理解、操作简单,但是对于不同的数据集,网格的大小很难确定[14]。本文经过多次尝试,以 15 × 11 进行划分最为理想,并根据统计得出出租车分布情况。

图 2 为出租车分布的热力图,图 2(a)为第 1 个时段内的上车分布情况,图 2(b)为第 2 个时段内的上车分布情况,同时图 2 也在一定程度上说明了不同时段上车点分布变化。

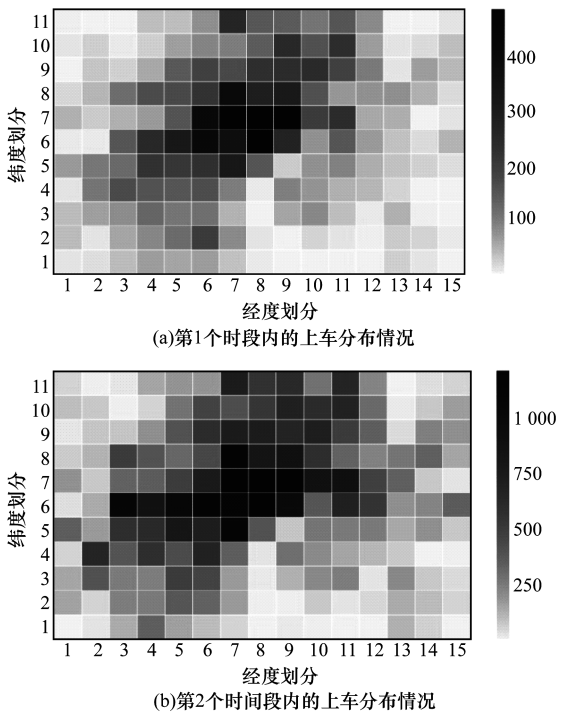

图 2　出租车分布热力图

本文根据分布情况在不同时段、不同范围内设置不同参数,以保证得到较好的热门区域结果。

1.3 出租车热门区域聚类发现

本文中所定义的热门区域包括热门上车区域及热门下车区域。对 5 个时段的出租车上车点、下车点进行聚类,聚类算法思想受到基于密度的聚类 DBSCAN 算法的启发,但是不同时段采用不同的参数(*eps* 和 *minpts*),分别得到在不同时段的热门区域集 P_{it}(热门上车区域)和 D_{it}(热门下车区域),其中,i 表示第 i 个热门区域集;t 表示时段。图 3 为第 2 个时段的热门区域集的聚类结果 P_{i2}(每个时段的热门区域集中的热门区域数量有差异)。由于不同时段的热门区域存在重叠情况,因此将 10 组不同结果进行合并,将所有不重叠的上车和下车热门区域合并,得到最终的 74 个出租车热门区域,如图 4 所示。

图 3 第 2 个时间段内热门上车区域的聚类结果

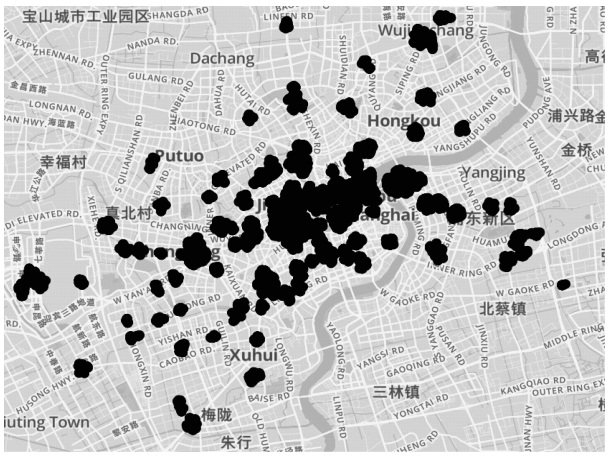

图 4 出租车热门区域

2 热门区域乘客出行特征发现

热门区域的乘客出行特征在一定程度上由该区域的功能决定。本文通过对热门区域内的乘客出行特征进行抽取,结合文本中广泛使用的 LDA 主题模型,得出热门区域的潜在功能主题。

2.1 LDA 主题模型

主题模型是一个生成性的贝叶斯网络,被广泛应用于发现大量文档中隐含主题的分布[15-16]。在 LDA 模型中,本文认为文档由隐含主题按照一定比例混合而成,而每个词都由一个主题生成并服从主题在词汇上的概率分布。

2.2 基于 LDA 的热门区域乘客出行特征发现

将每一个热门区域视作一个文档,某一个热门区域的区域功能视作文档的主题,每一个热门区域的客流模式相当于文档的单词。如同文档的单词是由文档的主题生成并为文档的主题推导提供帮助,热门区域的乘客出行特征在一定程度上由该区域的功能决定,并反映了该区域的功能。

区域功能发现与文档主题发现类比关系如表 2 所示。设 M 表示热门区域“文档”数目,N 为根据乘客出行规律聚类的客流峰段数目,则定义单词 $Word=(pattern,M_i,T_j)$,其中,*pattern* 取值为 *Pickup*(上车)或 *Dropoff*(下车);M_i 表示第 i 个热门区域;T_j 表示第 j 个时段。例如,假设热门区域 m 文档有 d 个单词($Dropoff,M_i,T_j$),表示在 T_j 时段有 d 个乘客离开 M_i 抵达 m;有 p 个单词($Pickup,M_i,T_j$),表示在 T_j 时段有 p 个乘客离开站点 m 抵达 M_i。最终对于某一区域 M_i 而言,得到 2 个“词频”矩阵,分别为 ***MatrixP*** 与 ***MatrixD***,矩阵大小为 $M\times N$,代表上车(或离开,*Pickup*)与下车(或抵达,*Dropoff*)情况统计,以此作为 LDA 输入的“单词”。

表 2 区域功能发现与文档主题发现类比

区域功能发现	文档主题发现
热门区域	文档
区域功能	文档的主题
客流模式	单词

由此,本文得到基于 LDA 的功能区域聚类模型的概率图表示[15],如图 5 所示。

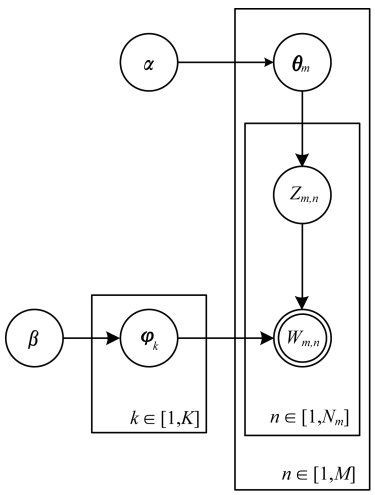

图 5 LDA 概率图表示

在图 5 中,K 表示功能区域数目;M 表示热门区域个数,每个热门区域有 N_m 个单词;$W_{m,n}$ 表示热门区域对应的每个单词;$Z_{m,n}$ 表示生成该单词的热门区域功能;$\boldsymbol{\theta}_m$ 为一个 K 维向量,对应热门区域 m 的功能描述生成 $Z_{m,n}$ 的多项式分布;$\boldsymbol{\varphi}_k$ 为一个向量,长度为单词种类数,$\boldsymbol{\varphi}_k$ 符从狄利克雷分布,分布参数为 β;$\boldsymbol{\varphi}_{Z_{m,n}}$ 是生成 $W_{m,n}$ 的多项式分布;$\boldsymbol{\theta}_m$ 服从狄利克雷分布,分布参数为 α。

LDA 模型生成过程如下[15]:

1)为每一个主题 k,根据狄利克雷分布及参数 β 生成 $\boldsymbol{\varphi}_k$。

2)为每一个热门区域 m,根据狄利克雷分布及参数 α 生成 $\boldsymbol{\theta}_m$。

3)为热门区域 m 的第 d 个单词 $W_{m,d}$,根据多项式分布及参数 $\boldsymbol{\theta}_m$,生成单词对应热门区域功能 $Z_{m,d}$,根据多项式分布及参数 $\boldsymbol{\varphi}_{Z_{m,d}}$,生成 $W_{m,d}$。

2.3　LDA 模型推导

本文采用吉布斯采样估计模型的参数,主要采样隐变量为 z。对于条件概率 $P(z_{i,d}=k\mid z_{\neg i,d},w)$,其中 $z_{\neg i,d}$ 代表除了 $w_{i,d}$ 以外所有词的主题指派,由吉布斯采样公式可得:

$$P(z_{i,d}=k\mid z_{\neg i,d},w)$$

$$\propto P(z_{i,d}=k,w_{i,d}=t\mid z_{\neg i,d},w_{\neg i,d})=\hat{\theta}_{ik}\cdot\hat{\varphi}_{kt}$$

其中,$\hat{\theta}_{ik}$,$\hat{\varphi}_{kt}$ 代表文档-主题与主题-单词的后验分布,利用 Dirichlet 分布与多项式分布的共轭,得到贝叶斯参数估计:

$$\hat{\theta}_{ik}=\frac{n^k_{i,\neg d}+\alpha_k}{\sum_{k=1}^{K}(n^k_{i,\neg d}+\alpha_k)}$$

$$\hat{\varphi}_{ik}=\frac{n^t_{k,\neg i,d}+\beta_t}{\sum_{t=1}^{V}(n^k_{k,\neg i,d}+\beta_t)}$$

随机初始化概率矩阵 $\boldsymbol{\theta}$,$\boldsymbol{\varphi}$,采样文档中词的主题指派并重新估计这 2 个矩阵,迭代直到收敛。据此可以得到每个区域对应的区域(文档)与功能(主题)的分布以及功能与客流模式(单词)的分布。之后,以分布情况为特征进行区域聚类,解释区域功能。

3　实验与结果分析

根据本文提出的基于 LDA 主题模型的热门区域乘客出行特征聚类模型的原理,每个热门区域的主要划分依据就是区域内的乘客出行模式,而相同的出行模式在一定程度上说明了该区域的功能。

3.1　实验数据

本文实验中使用的数据是 2015 年 4 月份共 20 000 辆强生出租车的行驶数据(数据来自 SODA "游族杯"上海开放数据创新应用大赛中上海强生出租车行车数据,比赛网址 http://soda.datashanghai.gov.cn/),这 20 000 辆出租车每隔 20 s 提供一次当前状态数据,其数据构成如下:{车 ID,载客状态,接收时间,GPS 测定时间,经度,纬度,速度},其中总共包含 30 亿条 GPS 数据。将数据按车辆 ID 进行划分并按时间进行排列,车辆状态由 0 变为 1 的 2 条相邻记录被认为是有乘客上车,而由 1 变为 0 的 2 条相邻记录被认为有乘客下车。从所有 GPS 数据中筛选出车辆状态变化的记录,得到乘客上车点与下车点的记录各 1 200 万条。

3.2　区域划分

区域划分部分使用 Spectral Clustering 聚类算法对区域进行划分,需要确定聚类的区域数目 S。经过多次参数选择,观察结果后发现 $S=6$ 时区域划分结果最为理想。区域划分结果如图 6 所示。可以发现,分类相同的热门区域点在地理位置上也聚集在一起,而不是零散地分布。这是因为上海轨道交通十分发达,考虑到交通成本,打车范围不会很大,所以同一类热门区域相对集中。同时,这也说明城市城区整体上也是按功能进行划分。

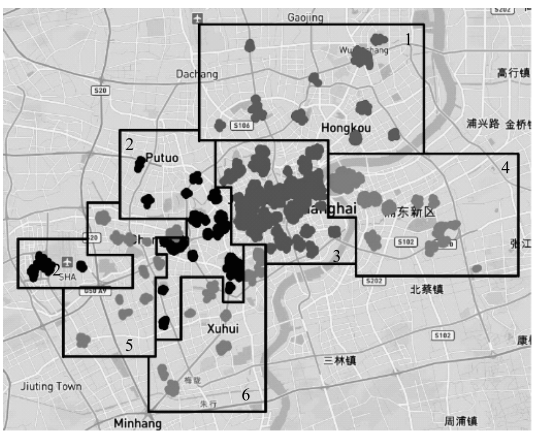

图 6　区域划分示意图

3.3　区域基本概况

聚类得到的 6 类功能区域由于其功能特点不同,因此其功能决定的乘客出行模式也不同。在表 3 中分别为 6 类功能区域在不同时间段内的上车、下车人数进行统计。通过观察表中数据发现在不同区域、不同时段内,乘车人数有很大差别。虽然受到一些如区域面积、时间跨度长短等因素影响,但在一定程度上还是说明了各区域功能不同的特点。如区域 4 的第 2 个时段内,上车人数 429,下车人数 653,而其他时段内并没有这么明显的差距,很可能区域 4 具有工作区域属性,吸引了更多的客流(即到达该区域客流为多);区域 5 的第 2 个时段内,上车人数 190,下车人数 85,那么区域 5 很可能具有居住区域属性,更多的客流流失(即上车离开该区域客流为多)。

表 3 各区域分时段上下车人数

区域	01:00:00—05:59:59		06:00:00—09:59:59		10:00:00—16:59:59		17:00:00—19:59:59		20:00:00—00:59:59	
	上车人数	下车人数	上车人数	下车人数	上车人数	下车人数	上车人数	下车人数	上车人数	下车人数
1	52	57	334	233	795	682	188	265	319	357
2	202	242	750	772	3 194	3 106	1 058	989	1 349	1 167
3	1 215	1 217	4 169	3 512	14 964	15 145	4 323	4 595	5 788	6 227
4	62	71	429	653	2 454	2 507	674	565	571	469
5	48	43	190	85	549	489	181	232	289	309
6	36	46	299	286	813	872	186	153	165	250

表 4 中是各区域的基本信息统计，如面积、热门区域个数等，其中总面积计算是找出簇内上下左右 4 个顶点，根据经纬度计算簇的长宽，近似为长方形计算面积，由于面积表示相对大小，因此无实际单位。

表 4 各区域基本信息统计

区域	总面积	热门区域个数
1	3.09	10
2	5.54	18
3	22.53	14
4	4.26	12
5	1.68	11
6	2.06	9

3.4 整体与区域客流分布对比

笔者发现在所有热门区域间的客流量和整体数据在分布上存在差异（图 1 所示为整体数据分布），如图 7 所示，整体数据按时段统计后与区域 2 的上车数据进行对比。由于区域数据与整体数据在数量上差异较大，因此图中对数量进行线性函数归一化处理，映射到[0,1]区间，从而比较区域与整体上的分布情况。

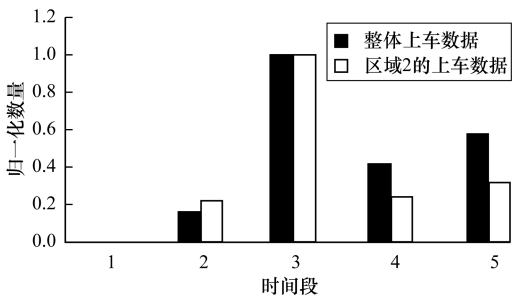

图 7 整体与区域客流数据分布情况

乘客出行模式包括热门区域间、热门区域与非热门区域间（例如热门区域内上车、非热门区域内下车）、非热门区域间的移动。所以，针对这种分布差异的问题，认为相比于 1,3 时间段，4,5 时间段的客流移动分布更为广泛，许多客流移动并没有局限于热门区域间的移动。例如，时间段 5 为娱乐休闲集中的时段（20:00:00—00:59:59），人们的娱乐活动丰富，活动地点分布广泛，所以，客流的分布也相对广泛，并没有集中于热门区域间的移动。同时由于热门区域人口密集，在高峰时段会加剧区域内交通拥堵，导致乘客成功打车的数量降低，因此客流移动更多分布在热门区域与非热门区域间、非热门区域间的移动。

在分析图 7 的过程中发现，与其他交通客流数据（如轨道交通、公交车）不同，在早晚高峰时段，出租车的客流量未达到最高峰，相反在 10:00:00—16:59:59时段内客流量达到最高峰。笔者认为出现这一现象的原因有：

1）由于上海的轨道交通十分发达，因此出租车作为城市客运交通的一部分，扮演的是常规公共交通的重要补充这一角色。

2）同时早晚高峰严重的交通拥堵也使得出租车出行成为次要选择，有效的乘车次数也相对降低。

3）在地铁和公交系统停运期间，出租车成为主要的客运交通工具。

3.5 不同区域客流分布对比

热门区域自身的变化趋势在一定程度上反映了该区域的功能。由于各区域的基本属性不一致，因此在进行线性函数归一化处理后将 6 类区域进行对比，如图 8 所示，其中，图 8(a)为上车情况；图 8(b)为下车情况。

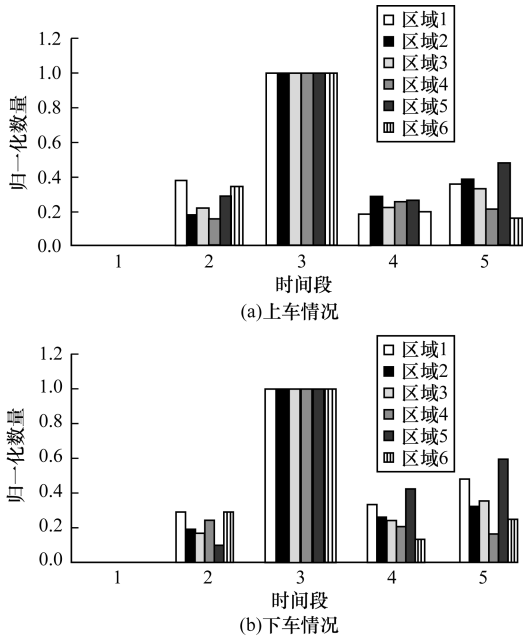

图 8 6 类区域客流数据分布情况

利用图 8 中展示的每一类区域各自的变化趋势可以发现一些区域的自身特点。例如,区域 5 在时间段 2 和时间段 5 的下车人数有很大差异;区域 1 在早高峰的上车人数明显大于晚高峰的上车人数;区域 3 的上车、下车情况分布基本一致。

3.6　区域功能总结

结合实际情况以及结果分析,对 6 类出租车热门区域进行功能总结,找出了其成为出租车热门区域的原因:

1)第 1 类区域以医院为主。在这一区域中大多数的热门区域都为医院,如上海新华医院、同济大学第十附属医院、同济医院、岳阳医院等。该区域内的乘客多在白天活动,而晚上相对要少,也符合人们去医院就医的生活习惯。

2)第 2 类区域以生活广场为主,其中包括上海悦达 889 广场、新世纪广场、金虹桥广场、百联国际广场等。作为局部的休闲娱乐场所,在工作时段,上下车人数分布均匀,无明显差异;在 20∶00∶00—01∶00∶00 时段,上车人数明显多于下车人数;在 01∶00∶00—06∶00∶00 时段,该区域的上下车人数要明显比区域 1、区域 4、区域 5、区域 6 多。这些乘客出行模式都符合休闲娱乐场所的客流特点。

3)第 3 类区域以地铁站为主。这类区域主要分布在市中心区域,地铁站点附近覆盖了各种属性的地点,如办公、居住、休闲娱乐等。所以,热门区域也主要集中在地铁站附近。该区域的乘车人数也明显多于其他区域。

4)第 4 类区域集中在浦东新区,以办公区域为主。在早高峰时段下车人数明显大于上车人数,而晚高峰则相反。同时,该区域在工作时段和非工作时段乘车人数比例明显高于其他区域。这些乘客出行模式都符合办公区域特点。

5)第 5 类区域以小区为主。通过实际情况观察,虽然许多热门区域为地铁站,但与第 3 类区域不同,该区域内的地铁站周围多为生活居住区,而其他类型的热门区域也主要为生活居住区,如上海荣华东道与水城南路交叉口、仙霞路与安龙路交叉口等。

6)第 6 类区域同样以医院为主。因为地理位置的不同,所以可能会对乘客出行模式产生影响,如上海复旦大学附属中山医院、华龙医院、复旦大学附属儿童医院等。

第 1 类和第 6 类区域经过实际分析发现虽然都是医院为主,但是它们的客流模式并不相同,这主要是因为这 2 个区域分别位于上海东北方向(第 1 类)和上海西南方向(第 6 类),区域的差异导致客流的模式存在差异。

功能区域中医院区域的发现在其他相关研究中很少发现。出租车具有方便快捷但价格相对较高的鲜明特点,而人们看病就医时总是希望尽快地将病人送到医院,不会计较交通费用问题,两者的特点是相吻合的。所以,出租车的热门区域中有相当一部分是医院。

在进行区域功能总结的过程中不难发现,有许多区域内的热门乘车区域在地铁站附近,这也证明了城市交通中存在的“最后一公里”问题[17]。

3.7　效果评估

将上文得出的区域功能发现结果与实际情况作对比,本文使用 RI(Rand Index)聚类评价指标对结果进行评估[18],具体计算如下:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

其中,TP 表示同一类的热门区域被分到同一个簇;TN 表示不同类的热门区域被分到不同簇;FP 表示不同类的热门区域被分到同一个簇;FN 表示同一类的热门区域被分到不同簇。通过计算得出聚类结果的 RI 为 84.3%,说明本文提出方法能够有效发现热门区域的功能特点。

4　结束语

本文以出租车客流数据为基础发现了热门区域,然后基于 LDA 主题模型提出对热门区域进行聚类的模型,结果体现了区域的功能和不同时段内区域与客流的关系,并对这些区域成为热门区域的背后原因进行分析,对于提高城市规划、乘客寻车和出租车寻客的资源利用率具有指导作用。下一步将在区域划分方式上进行改进,结合 POI 等数据,更好地解决此类具有时空特性的交通数据聚类问题。

参考文献

[1] Lee J, Shin I, Park G L. Analysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation[C]// Proceedings of International Conference on Networked Computing and Advanced Information Management. Washington D. C., USA: IEEE Computer Society, 2008: 199-204.

[2] Chang H W, Tai Y C, Hsu Y J. Context-aware Taxi Demand Hotspots Prediction[J]. International Journal of Business Intelligence & Data Mining, 2010, 5(1): 3-18.

[3] Palma A T, Bogorny V, Kuijpers B, et al. A Clustering-based Approach for Discovering Interesting Places in Trajectories[C]//Proceedings of 2008 ACM Symposium on Applied Computing. New York, USA: ACM, 2008: 863-868.

[4] Yue Yang, Zhuang Yan, Li Qingquan, et al. Mining Time-dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data[C]//Proceedings of the 17th International Conference on Geoinformatics. Washington D. C., USA: IEEE Press, 2009.

[5] Veloso M, Phithakkitnukoon S, Bento C. Sensing Urban

Mobility with Taxi Flow[C]// Proceedings of International Workshop on Location-based Social Networks. New York, USA: ACM Press, 2011: 41-44.

[6] Phithakkitnukoon S, Veloso M, Bento C, et al. Taxi-aware Map: Identifying and Predicting Vacant Taxis in the City[C]//Proceedings of the 1st International Joint Conference on Ambient Intelligence. Berlin, Germany: Springer-Verlag, 2010: 86-95.

[7] Pan Gang, Qi Guande, Wu Zhaohui, et al. Land-use Classification Using Taxi GPS Traces [J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1): 113-123.

[8] Li Xiaolong, Pan Gang, Wu Zhaohui, et al. Prediction of Urban Human Mobility Using Large-scale Taxi Traces and Its Applications[J]. Frontiers of Computer Science, 2012, 6(1): 111-121.

[9] Yuan Jing, Zheng Yu, Xie Xing. Discovering Regions of Different Functions in a City Using Human Mobility and POIs[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2012: 186-194.

[10] 冷 彪, 赵文远. 基于客流数据的区域出行特征聚类[J]. 计算机研究与发展, 2014, 51(12): 2653-2662.

[11] Liu Chengkun, Qin Kun, Kang Chaogui. Exploring Time-dependent Traffic Congestion Patterns from Taxi Trajectory Data[C]//IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. Washington D. C., USA: IEEE Press, 2015: 39-44.

[12] Sander J, Ester M, Kriegel H P, et al. Density-based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications [J]. Data Mining & Knowledge Discovery, 1998, 2(2): 169-194.

[13] Ester M, Kriegel H P, Jiirg S, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 1996: 226-231.

[14] Castro P S, Zhang Daqing, Chen Chao, et al. From Taxi GPS Traces to Social and Community Dynamics: A Survey[J]. ACM Computing Surveys, 2014, 46(2): 1167-1182.

[15] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(2): 993-1022.

[16] Steyvers M, Griffiths T. Probabilistic Topic Models[J]. Handbook of Latent Semantic Analysis, 2007, 427(7): 424-440.

[17] Balcik B, Beamon B M, Smilowitz K. Last Mile Distribution in Humanitarian Relief [J]. Journal of Intelligent Transportation Systems, 2008, 1818(2): 51-63.

[18] Rand W M. Objective Criteria for the Evaluation of Clustering Methods [J]. Journal of the American Statistical Association, 1971, 66(336): 846-850.

编辑 陆燕菲

(上接第 15 页)

[16] 朱金奇, 马春梅, 刘 明, 等. 车载自组织网络中基于停车骨干网络的数据传输[J]. 软件学报, 2016, 27(2): 432-450.

[17] 刘冰艺, 吴黎兵, 贾东耀, 等. 基于移动云服务的车联网数据上传策略[J]. 计算机研究与发展, 2016, 53(4): 811-823.

[18] 冯 成, 李志军, 姜守旭, 等. 无线移动感知网络上的数据聚集传输规划[J]. 计算机学报, 2015, 38(3): 685-700.

[19] You Kun, Tang Bin, Qian Zhuzhong, et al. QoS-aware Placement of Stream Processing Service [J]. Journal of Supercomputing, 2013, 64(3): 919-941.

[20] Rochman Y, Levy H, Brosh E. Resource Placement and Assignment in Distributed Network Topologies [C]// Proceedings of IEEE INFOCOM'13. Washington D. C., USA: IEEE Press, 2013: 1914-1922.

[21] Rochman Y, Levy H, Brosh E. Efficient Resource Placement in Cloud Computing and Network Applications[J]. ACM SIGMETRICS Performance Evaluation Review, 2014, 42(2): 49-51.

编辑 金胡考