

基于客流数据的区域出行特征聚类

冷彪 赵文远
(北京航空航天大学计算机学院 北京 100191)
(lengbiao@buaa.edu.cn)

Region Ridership Characteristic Clustering Using Passenger Flow Data

Leng Biao and Zhao Wenyuan
(School of Computer Science and Engineering, Beihang University, Beijing 100191)

Abstract Region function is an integral part of urban planning. The extraction and mining of ridership characteristic can be regarded as data support of region function recognition. The advance of intelligent transportation technology in metro system enables the collection of spatial-temporal passenger flow data, which conveys human mobility and indicates the similarity between metro stations, also closely related to the region function during different periods. This paper discusses the ridership characteristic clustering using passenger trip pattern and metro station flow pattern extracted from metro passenger flow data. Firstly, we identify the passenger flow centrality and station tide flow from passenger trip pattern and metro station flow pattern, which imply the region function of metro stations. Secondly, by discovering the similarity between region cluster and text analysis, we take advantage of the classical probabilistic graphical model and propose a novel LDA-based region ridership characteristic clustering model, allocating metro stations with similar ridership characteristic into the same region. Thirdly, the experimental results show the passenger flow relationship among regions and recognize the region functions during different periods. The analysis of clustering results gives us a good understanding of how passenger flow circulates during different periods and may enables many valuable services like network design and crowd evacuation.

Key words ridership characteristic clustering; human mobility; probabilistic graphical model; metro passenger flow; data mining

摘要 区域功能发现对完善城市规划有着重要的指导意义. 区域居民的出行特征提取与发掘可以作为建立模型分析区域功能的数据支撑. 随着智能交通技术在轨道交通系统的应用, 大量蕴含行人移动性和出行目的地信息的客流数据被采集得到, 发现客流数据与地铁站相关区域功能有紧密联系. 从地铁客流数据中提取出乘客出行模式和地铁站客流模式, 并以此为基础建立概率图模型, 实现了区域出行特征聚类. 首先, 以地铁客流数据为基础提取了乘客出行模式和地铁站客流模式, 发现地铁站客流集中性和潮汐性的特性, 能在一定程度上反映地铁的区域功能. 然后, 采用了文本分析领域经典的概率图模型, 建立基于潜在狄利克雷分配(latent Dirichlet allocation, LDA)主题模型的地铁客流出行特征聚类模型, 将具有出行规律相似性的地铁站聚类在一起. 最后, 通过分析聚类实验结果, 发现在不同客流峰段内的区域功能和相互客流关系.

关键词 出行特征聚类;人类移动性;概率图模型;地铁客流;数据挖掘

中图法分类号 TP391

城市间不同区域居民的出行数据蕴含了行人移动性和出行目的信息,能够帮助研究人员识别出区域间的交通客流关系和区域功能,可以为改善交通、优化城区规划提供宝贵的指导信息.

区域功能发现对完善城市规划有重要的指导意义^[1].早期的城市区域功能发现主要依赖于卫星或遥感图像^[2].但是,基于图像的分析方法需要人工标注区域的地理位置,对于处理大量数据效率较低^[3].随着智能交通系统的不断发展,出租车轨迹数据逐渐应用到人类移动性分析和区域功能划分领域. Liu 等人实时监控城区居民移动模式,建立了实时城区街景系统^[4]. Goh 等人实时采集和监控交通信息,并建立了在线地图匹配系统^[5]. Liu 等人通过对交通数据流进行分析,发现了数据流的时间-空间交互关系^[6]. Guc 提出了一种将轨迹数据与区域功能分析结合的新思路,对出租车的轨迹数据进行挖掘整理,并标记轨迹数据的语义^[7]. Yuan 等人根据出租车轨迹数据和兴趣点(points of interests, POIS)数据建立了基于潜在狄利克雷分配(latent Dirichlet allocation, LDA)的区域功能非监督聚类模型^[8]. Pan 等人提取了区域以及区域相关的轨迹数据特征,通过手工标注部分区域功能,建立基于支持向量机(support vector machine, SVM)的分类模型对区域的功能进行划分,还应用轨迹数据对出行进行了乘客预测^[9-10]. Sheng 等人提出了 POIS 数据的参考距离特征(reference distance feature),用来表征每一区域对应兴趣点(interests)的相互距离和排列模式,并建立了应用 Reference Distance Feature 对区域功能进行聚类的空间向量场(spatial vector space)模型^[11]. Xue 等人采用 Markov 随机域(Markov random field)对乘客出行的可能去向进行建模表示,实现了对乘客出行目的的预测^[12]. Zheng 等人分别根据 GPS 轨迹数据和用户在不同地点的签到信息数据挖掘提炼了位置与位置之间的关系^[13-14].此外,少数研究学者利用地铁客流数据对城市区域功能进行分析. Chen 等人以纽约地铁系统为研究目标,分析了纽约地铁乘客的出行模式^[15]. Leng 等人针对地铁业务提出了基于概率树的客流预测方法^[16].

地铁客流数据与地铁站附近区域功能有密切的关系,为区域功能发现提供了具有时间和空间双重属性的指导信息.时间属性指的是乘客在一天中某一个时段抵达或者离开某一个地铁站.空间属性指

的是乘客从起始站出发抵达目的站.

本文采用地铁客流数据,建立了基于 LDA 的区域功能聚类模型,对区域客流特征进行聚类分析.本文的主要贡献如下:

- 1) 以地铁客流数据为基础提取了乘客出行模式和地铁站客流模式;
- 2) 建立了基于 LDA 主题模型的区域出行特征聚类模型,将具有类似客流时空特性的地铁站聚类在一起;
- 3) 分析聚类结果,发现不同客流峰段内的区域功能和相互客流关系.

1 地铁客流数据描述与分析

1.1 源数据描述

本文使用的地铁客流数据来源于北京市交通委和北京市轨道交通指挥中心.截止 2013 年 5 月,北京市轨道交通路网共有 17 条运营线路、270 座运营车站,运营线路总长 456 km.

本文采集了北京市轨道交通路网 2013 年 6 月所有工作日的客流数据,平均每天大约 400 多万条出行记录.每条出行记录包含以下信息:

- 1) 乘客的起始站点 ID,确定地铁站点的唯一标识;
- 2) 乘客在起始站点进站刷卡时刻,格式为“YYYY-MM-DD-HH-MM”,分钟精确到 5 min;
- 3) 乘客的目的站点 ID,确定地铁站点的唯一标识.

1.2 OD 记录定义

单个乘客的一次出行产生一条起始点(origin-destination, OD)记录,本文定义为 R . OD 记录 R 由 3 部分组成:乘客出行的起始站点 $R.oStation$ 、乘客出行的目的站点 $R.dStation$ 以及乘客在起始站点的刷卡时段 $R.oTime$. 一条记录 $R:(R.oStation, R.dStation, R.oTime)$ 表示存在一个乘客在时刻 $R.oTime$ 刷卡进站,由地铁站 $R.oStation$ 去往地铁站 $R.dStation$. 根据乘客出行的 OD 记录,可以提取乘客出行模式(passenger trip pattern).

1.3 乘客出行模式

乘客出行模式反映了乘客的移动性信息和出行目的,可以通过 OD 记录提取.

乘客出行模式本文定义为 T , 由 3 部分组成: 出行的起始站点 $T.oStation$ 、出行的目的站点 $T.dStation$ 和出行的客流峰段 $T.oPeriod$.

乘客的出行模式不但反映了乘客的出行选择, 同时能够在一定程度上体现乘客选择的起始站点 $T.oStation$ 和目的站点 $T.dStation$ 在客流峰段 $T.oPeriod$ 的功能, 所以本文在乘客出行模式的基础上针对每个地铁站提取了地铁站客流模式 (metro station flow pattern).

1.4 地铁站客流模式

地铁站客流模式反映了地铁站在客流峰段内的客流规律以及地铁站与其他地铁站的客流关系. 对于一个地铁站 s , 本文定义其客流模式为 F_s . F_s 由 2 个 $S \times N$ 矩阵组成, 分别是到站客流矩阵 M_A 和离站客流矩阵 M_D , 其中 S 是路网中地铁站的数目, N 为根据乘客出行规律划分的客流峰段数目.

在乘客出行模式的基础上为提取了地铁站客流模式, 对于地铁站 s , 根据乘客出行模式集合 $\{T\}$, 其到站客流矩阵 M_A 的 (i, j) 单元的值 $M_A(i, j) = \|\{T=(x, y, z) | x=i, y=s, z=j\}\|$, 表示在峰段 j , 由 $M_A(i, j)$ 数量的乘客离开地铁站 i 抵达地铁站 s .

同理, 离站客流矩阵 M_D 的 (i, j) 单元的值 $M_D(i, j) = \|\{T=(x, y, z) | x=s, y=i, z=j\}\|$ 表示在峰段 j , 由 $M_D(i, j)$ 数量的乘客离开地铁站 s 抵达地铁站 i .

1.5 地铁客流峰段划分

地铁客流峰段是北京市交通委、北京市轨道交通指挥中心根据乘客出行规律, 按照轨道交通线路一天的运营时间以及客流量大小划分的时间段, 划分规则如表 1 所示:

Table 1 Period Partition Rule

表 1 客流峰段划分表

Period	Start Time	End Time
Morning Flat	05:00:00	06:59:59
Morning Peak	07:00:00	08:59:59
Noon Flat	09:00:00	16:59:59
Evening Peak	17:00:00	19:29:59
Evening Flat	19:30:00	23:29:59

一般来说, 在同一峰段乘客出行的目的基本一致, 例如工作日的早高峰阶段大量的客流是由于乘客上班产生的通勤客流, 此时客流的基本流动方向是由居住区附近的地铁站去往工作区附近地铁站或者交通枢纽附近的地铁站.

1.6 乘客出行数据统计与分析

本文对地铁客流数据进行了统计分析, 提取了乘客出行模式和地铁站客流模式, 发现了以下 2 种现象.

1.6.1 站点客流集中性

如图 1 和图 2 所示, 存在一类站点的进站客流和出站客流量相当大, 说明了轨道交通客流很大程度上集中在这一类站点. 在 228 个地铁站点中, 有 20 个站点吸引了 27.1% 的进站客流, 同时又有 20 个站点释放了 29.5% 的出站客流. 并且, 出站客流的站点集中程度要强于进站客流的站点集中程度.

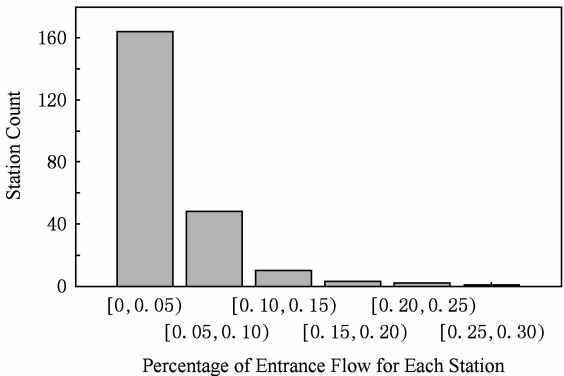


Fig. 1 Histogram for percentage of entrance passenger flow.

图 1 进站客流比例柱状统计图

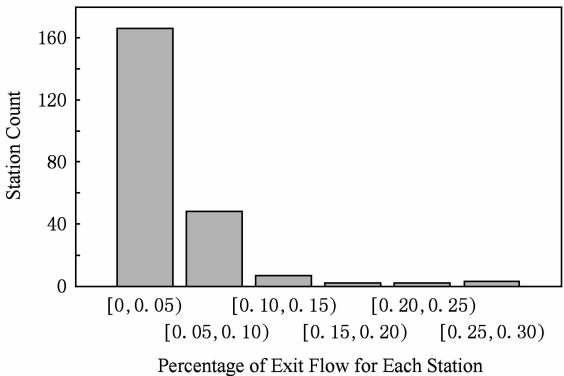


Fig. 2 Histogram for percentage of exit passenger flow.

图 2 出站客流比例柱状统计图

这类客流集中性的站点主要代表有交通枢纽的北京站, 工作、娱乐中心的西单以及核心居住区天通苑、苹果园等. 观察这类站点的客流模式, 会发现它们的到站客流矩阵和离站客流矩阵均存在某个时段值比较大. 观察其他站点的客流模式会发现它们的到站客流矩阵和离站客流矩阵中对应具有客流集中性的站点的矩阵值也相对较大.

1.6.2 站点客流潮汐性

部分站点的进站客流和出站客流呈现出一定程度的潮汐特性. 在一天的某几个峰段, 站点的进站客流比较大, 站点的出站客流比较小; 对应地在一天的另外几个峰段, 站点的出站客流比较小, 站点的进站客流比较大.

如图 3 和图 4 所示, 对于核心居住区苹果园等, 站点客流潮汐性比较强. 而对于工作、娱乐中心西单, 站点客流的潮汐性就比较弱.

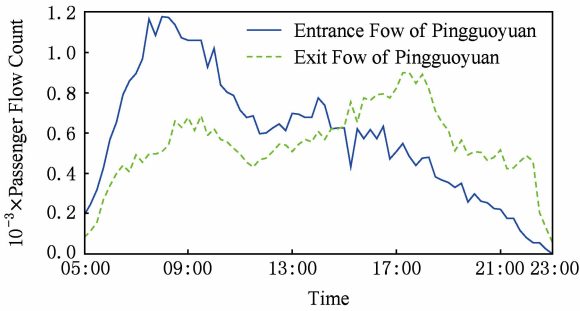


Fig. 3 Entrance and exit passenger flow of Pingguoyuan.
图 3 苹果园进出站客流潮汐性统计图

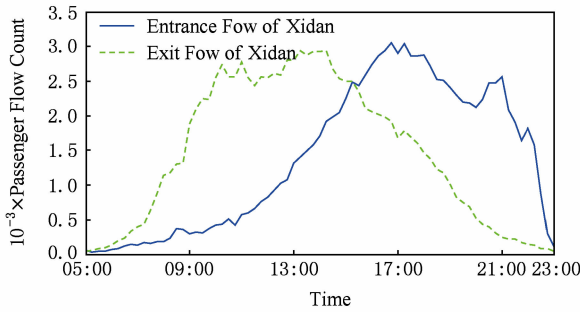


Fig. 4 Entrance and exit passenger flow of Xidan.
图 4 西单进出站客流潮汐性统计图

观察核心居住区苹果园的客流模式, 发现离站客流矩阵在早高峰对应列的值较大, 晚高峰对应列的值较小, 到站客流矩阵在早高峰对应列的值较小, 晚高峰对应列的值较大. 而作为工作、娱乐中心西单的客流模式, 也存在一定的客流潮汐性, 但进出站客流高峰到来顺序与居住区站点恰好相反.

2 基于 LDA 的地铁站出行特征聚类模型

根据对地铁客流数据的描述与分析, 本文提取了乘客出行模式和站点客流模式, 发现了站点客流集中性和潮汐性的现象. 同时发现客流集中性和潮汐性的强弱与作用规律, 能在一定程度上揭示站点

的区域功能, 而集中性和潮汐性又能被客流的出行模式和站点的客流模式量化反映. 所以, 本文提出了基于 LDA 的地铁站出行特征聚类模型, 应用乘客出行模式与站点客流模式对地铁站按出行特征进行区域聚类, 通过分析聚类结果识别出区域的功能和客流关系.

2.1 LDA 主题模型简介

LDA 主题模型是一个生成性的 3 层贝叶斯网络, 已经被广泛应用到文档分类、图像视频标注等领域^[17-18]. 根据 LDA 模型理论, 文档库中的每一个文档都可以看作是由多个主题混合产生的, 文档的每一个单词都是来自一个主题. 当给定一个文档的所有单词的情况下, 可以通过数学推导得到文档对应的主题分布. 通常, 对 LDA 模型进行推导有 Gibbs sampling 和 EM 两大算法.

2.2 基于 LDA 的出行特征聚类

本文将轨道交通相关数据与采用 LDA 模型进行文本分类的原始数据进行了比较. 将每一个地铁站视作一个文档, 地铁站对应的区域功能视作文档的主题, 则对于每一个地铁站, 其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, 地铁站的客流模式在一定程度上由地铁站的区域功能决定并反映出地铁站的区域功能.

根据轨道交通客流的意义将地铁站客流模式转化为地铁站“文档”对应的单词. 设 S 表示地铁站“文档”的数目, N 为根据乘客出行规律聚类的客流峰段数目. 则定义单词 $Word = (direction, S_i, T_j)$, 其中 $direction$ 取值为 Arriving 或 Departing, S_i 表示第 i 个地铁站, T_j 表示第 j 个客流峰段.

假设地铁站 s 文档有 m 个单词 (Arriving, S_i, T_j), 表示在客流峰段 T_j , 有 m 个乘客离开站点 S_i 抵达站点 s ; 有 n 个单词 (Departing, S_i, T_j), 表示在客流峰段 T_j , 有 n 个乘客离开站点 s 抵达站点 S_i . 对于地铁站 s , 其客流模式 F_s 向对应的单词转换如图 5 所示.

由此, 本文得到了基于 LDA 的功能区域聚类模型的概率图表示, 如图 6 所示.

图 6 中 K 表示功能区域数目, S 表示地铁站的个数, 每个地铁站有 N 个单词. $W_{s,n}$ 表示地铁站对应的每个单词, $Z_{s,n}$ 表示生成该单词的地铁站功能. θ_s 为一个 K 维向量, 对应地铁站 s 的功能描述是生成 $Z_{s,n}$ 的多项式分布. β_k 为一个向量, 向量长度为单

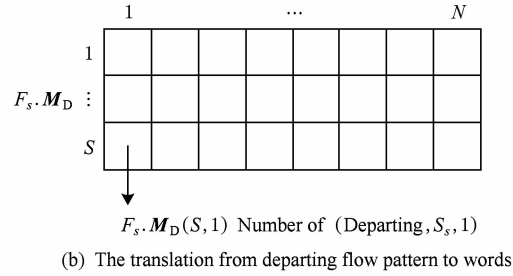
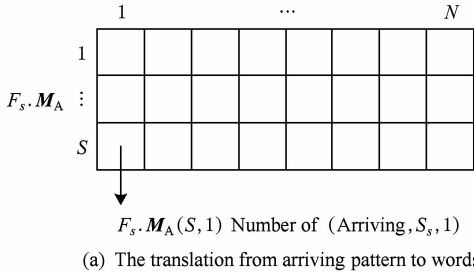


Fig. 5 The translation from flow pattern to words.

图 5 客流模式向对应的单词转换示意图

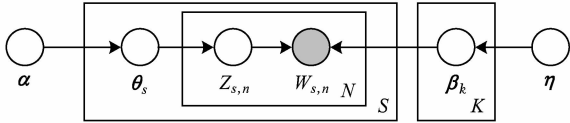


Fig. 6 The representation of clustering model.

图 6 聚类模型的概率图表示

词的种类数, $\beta_{Z_{s,n}}$ 是生成 $W_{s,n}$ 的多项式分布. $\theta_i (i = 1, \dots, S)$ 符合狄利克雷分布, 分布的参数为 $\alpha, \beta_i (i = 1, \dots, K)$ 符合狄利克雷分布, 分布的参数为 η .

LDA 模型的生成过程如下:

- 1) 为每一个主题 k , 根据狄利克雷分布以及参数 η , 生成 $\beta_k (k = 1, \dots, K)$;
- 2) 为每一个地铁站 s , 根据狄利克雷分布以及参数 α , 生成 $\theta_s (s = 1, \dots, S)$;
- 3) 为地铁站 s 的第 d 个单词 $W_{s,d}$, 根据多项式分布以及参数 θ_s , 生成单词对应地铁站功能 $Z_{s,d}$, 根据多项式分布以及参数 $\beta_{Z_{s,d}}$, 生成单词 $W_{s,d}$.

3 模型数学推导

根据原始数据提取出客流出行模式和地铁站客流模式, 采用 LDA 模型进行地铁站出行特征聚类. 模型推导符号定义说明由表 2 所示:

Table 2 Definition of Variables InCluster Model

表 2 聚类模型符号定义表

Variable	Definition
K	Number of Regions
S	Number of Stations
V	Kinds of Words
N_s	Words of Station s
$\theta_i (i = 1, \dots, S)$	Topic Distribution of Station i
$\beta_i (i = 1, \dots, K)$	Word Distribution over Word
α	Hyperparameter of $\theta_i (i = 1, \dots, S)$
η	Hyperparameter of $\beta_i (i = 1, \dots, K)$
$Z_{s,d}$	Region Assignment for $W_{s,d}$
$W_{s,d}$	d th Word for Station s

采用 Gibbs Sampling 对模型进行推导, 实际上就是计算条件概率分布 $P(Z_{s,d} | W, Z_{\neg s,d}, \alpha, \eta)$, 其中 W 表示所有文档的单词集合, $Z_{\neg s,d}$ 表示对单词 $W_{s,d}$ 以外的所有单词对应功能区域的赋值集合.

采用狄利克雷分布是多项式分布的共轭先验分布这一特性, 简化计算过程, 得到:

$$\begin{aligned} P(Z_{s,d} | W, Z_{\neg s,d}, \alpha, \eta) &= \\ \frac{P(W_{s,d}, Z_{s,d} | W_{\neg s,d}, Z_{\neg s,d}, \alpha, \eta)}{P(W_{s,d} | W_{\neg s,d}, Z_{\neg s,d}, \alpha, \eta)} &= \\ \frac{P(W, Z | \alpha, \eta)}{P(W, Z_{\neg s,d} | \alpha, \eta)} &= \\ \frac{P(W, Z | \alpha, \eta)}{P(W_{\neg s,d}, Z_{\neg s,d} | \alpha, \eta) P(W | \alpha, \eta)} &\propto \\ \frac{P(W, Z | \alpha, \eta)}{P(W_{\neg s,d}, Z_{\neg s,d} | \alpha, \eta)} &\propto \\ \frac{n_{Z_{s,d}}^{w_{s,d}} + \beta_{w_{s,d}} - 1}{\sum_{v=1}^V (n_{Z_{s,d}}^v + \beta_v) - 1} \times (n_{m^{Z_{s,d}}}^Z + \alpha - 1), \quad (1) \end{aligned}$$

其中, $n_{Z_{s,d}}^v$ 表示第 v 类单词中赋予区域功能 $Z_{s,d}$ 的单词总数, $n_{m^{Z_{s,d}}}^Z$ 表示在地铁站 s 文档中赋予区域功能 $Z_{s,d}$ 的单词总数.

同样, 采用狄利克雷先验和贝叶斯化简规则, 计算参数 $\theta_s (s = 1, \dots, S)$ 和 $\beta_k (k = 1, \dots, K)$:

$$\begin{aligned} P(\theta_s | Z_s, \alpha) &= \frac{P(\theta_s, Z_s | \alpha)}{P(Z_s | \alpha)} = \\ \frac{P(Z_s | \theta_s) \times P(\theta_s | \alpha)}{P(Z_s | \alpha)} &= \\ \frac{\prod_{n=1}^{N_s} P(Z_s | \theta_s) \times P(\theta_s | \alpha)}{P(Z_s | \alpha)} &= \\ \frac{1}{P(Z_s | \alpha)} \prod_{k=1}^K \theta_{s,k}^{\alpha_k} \times \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{s,k}^{\alpha_k - 1} &= \\ Dirchlet(\theta_s | n_s + \alpha) &\propto \prod_{k=1}^K \theta_{s,k} (k)^{n_s^k + \alpha_k - 1}, \quad (2) \end{aligned}$$

其中, n_s 是 K 维向量, 表示地铁站 s 对于 K 个功能

中每个功能的单词个数.

$$P(\boldsymbol{\beta}_k \mid Z, W, \boldsymbol{\eta}) = \frac{P(\boldsymbol{\beta}_k, Z_s, W \mid \boldsymbol{\eta})}{P(Z, W \mid \boldsymbol{\eta})}$$
$$\frac{\prod_{m=1, n=1}^{m=S, n=N_S} P(W_{m,n} \mid \boldsymbol{\beta}_k) P(\boldsymbol{\beta}_k \mid \boldsymbol{\eta})}{P(Z, W \mid \boldsymbol{\eta})} =$$
$$\frac{1}{P(Z, W \mid \boldsymbol{\eta})} \prod_{v=1}^V \boldsymbol{\beta}_{k,v}^v \times \frac{\Gamma(\sum_{v=1}^V \boldsymbol{\eta}_v)}{\prod_{v=1}^V \Gamma(\boldsymbol{\eta}_v)} \prod_{v=1}^V \boldsymbol{\beta}_{k,v}^{\boldsymbol{\eta}_v-1} =$$
$$\text{Dirchlet}(\boldsymbol{\beta}_k \mid \boldsymbol{n}_k + \boldsymbol{\eta}) \propto \prod_{v=1}^V \boldsymbol{\beta}_k(v)^{\boldsymbol{n}_k^v + \boldsymbol{\eta}_v - 1}, \quad (3)$$

其中, $\boldsymbol{n}_k$ 是 V 维向量, 表示 V 类单词每类对应功能 K 的单词数.

根据狄利克雷分布的期望^[19], 化简得到:

$$\boldsymbol{\beta}_k(v) = \frac{\boldsymbol{n}_k^v + \boldsymbol{\eta}_v}{\sum_{v=1}^V (\boldsymbol{n}_k^v + \boldsymbol{\eta}_v)}, \quad (4)$$

$$\boldsymbol{\theta}_s(k) = \frac{\boldsymbol{n}_s^k + \boldsymbol{\alpha}_k}{\sum_{k=1}^K (\boldsymbol{n}_s^k + \boldsymbol{\alpha}_k)}, \quad (5)$$

其中, $\boldsymbol{n}_k^v$ 表示第 v 类单词中赋予区域功能 k 的单词总数, $\boldsymbol{n}_s^k$ 表示在地铁站 s 文档中赋予区域功能 k 的单词总数.

对于地铁站 s , 对应的功能区域为 j , 使得 $\boldsymbol{\theta}_i(j) = \max(\boldsymbol{\theta}_i)$.

在实验中需要预先设定的参数为 $\boldsymbol{\alpha}, \boldsymbol{\eta}$ 和聚类的区域数目 K . 一般情况下, $\boldsymbol{\eta} = \{0.1\}_V$, 而 $\boldsymbol{\alpha} = \{50/K\}_K$. 对于聚类数目通常采用 *Perplexity* 来确定^[20]:

$$\text{Perplexity} = \exp \left\{ - \frac{\sum_{s=1}^N \sum_{d=1}^{N_s} \ln P(W_{s,d})}{\sum_{s=1}^N N_s} \right\}. \quad (6)$$

Perplexity 越小, 表示选择的初始参数和模型的训练结果对数据集的解释性能越好, 通过计算不同聚类数目的 *Perplexity*, 确定聚类数目为 7.

4 聚类结果分析

工作日和周末的地铁客流有比较大的差异, 首先工作日的早、晚高峰客流相对集中密集, 占全天客流的比例高; 另外工作日早晚高峰出行的乘客的出行目的也相对类似. 采用工作日的客流数据进行聚类实验, 更易于根据乘客的出行目的分析地铁站点在不同时段的区域功能.

采用 2013 年 6 月所有工作日的客流数据进行聚类分析, 聚类结果如图 7 所示:

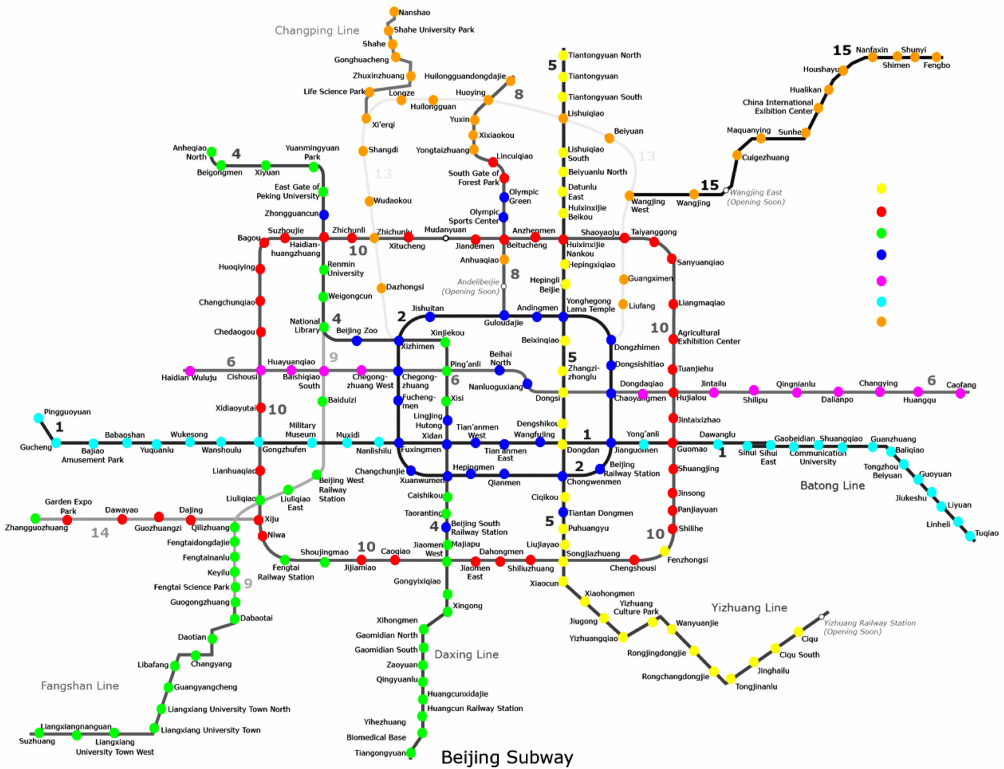


Fig. 7 Clustering result of data 2013.

图 7 聚类结果示意图

图 6 中区域 3 和区域 1 位于北京市城区的核心,区域 0,2,4,5,6 分别位于北京市城区的边缘,并且相同地铁站的站点大多属于同一区域。

根据聚类模型的原理,聚类模型为地铁站的每个单词指定一个对应的功能,那么,具有相似单词集合的地铁站就会被聚类在一起.地铁站的单词是根据地铁站的客流模式提取得到的,那么聚类在一起地铁站其单词集合相似,客流模式相似,也就是说在同一个客流峰段内,地铁站吸引的进站客流与地铁站释放的出站客流相似,即在同一个客流峰段内,区域内的地铁站功能基本一致。

按照区域聚类的结果对区域的客流模式进行分析,挖掘区域在不同时段的功能。

4.1 区域客流模式分析

根据聚类模型原理,对于一个站点,站点的单词种类越多,与该站点存在联系的站点就多;一个站点的某类单词的数目越多,该站点与对应站点的联系就越强。

对 7 个功能区域的地铁站单词种类数和单词数目进行统计,如表 3 所示:

Table 3 Statistics of Regions
表 3 功能区域站点单词统计

Region	Station Number	Word Kind per Station	Word Count per Station
0	31	1 756	547 020
1	42	1 807	590 340
2	50	1 663	517 900
3	31	1 914	1435 340
4	13	1 774	540 880
5	26	1 863	770 540
6	35	1 644	507 940

通过统计发现区域 3、区域 5 和区域 1 的平均单词种类数和站点平均单词数较大,区域 3、区域 5 和区域 1 可能是将所有区域联系起来的区域。但是现在下这个结论还为时过早,本文针对某些客流峰段内客流的流动规律进行了挖掘整理。

4.2 早高峰时段区域功能分析

早高峰客流峰段为 7:00—9:00,此时出行的客流大部分是市民上班产生的通勤客流,乘客的主要移动方向是从居住区去往工作区或者是换乘其他交通方式的工作枢纽.分析早高峰时段的乘客在区域间的流动,结果如图 8 所示。

图 8 中矩阵的每一行表示早高峰该区域进站客流去往其他区域的比例.发现每个区域的进站客流,

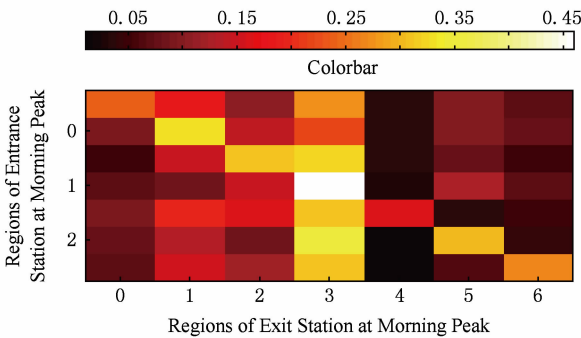


Fig. 8 Ridership statistics between regions at morning peak.

图 8 早高峰乘客区域出行比例统计图

主要去往区域 3、本身区域和区域 1.表明区域 3 和区域 1 在早高峰吸引了来自周围各个区域的大部分的进站客流.区域 1 和区域 3 在早高峰时段对应的功能为工作区,或者用来中转的交通枢纽.每个区域的进站客流仍有相当大的一部分去往本身区域,说明这些乘客居住和工作在同一区域,也就是说乘客倾向于居住在离工作近的地方,当然由于经济原因,工作在区域 3 和区域 1 的乘客往往不能居住在区域 3 和区域 1。

根据北京市轨道交通系统路网结构,采用了 Dijkstra 算法,根据乘客出行的起始站点和目的站点,计算出最短路径的站点数.统计了早高峰时段乘客出行平均站点数,结果如图 9 所示:

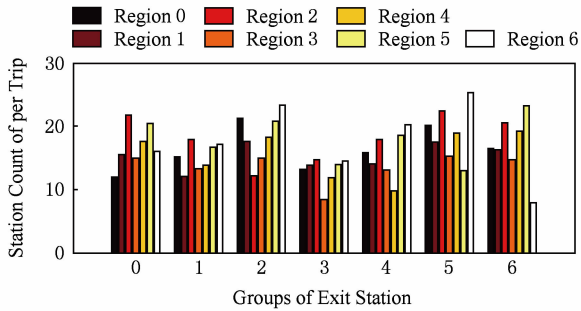


Fig. 9 Travel cost statistics at morning peak.

图 9 早高峰乘客出行平均站点数

通过统计发现区域内部旅行需要经过的站点数目是最少的,这一定程度上证明了居住和工作在同一区域使得早高峰上班花费的时间更短。

通过观察区域的地理分布,发现区域 0 和区域 6、区域 5 和区域 2 之间地理位置接近,可是区域 0 和区域 6 之间的客流少,区域 5 和区域 2 之间的客流比较少.这一方面是因为这 4 个区域相比较核心区域 3 和核心区域 1 并不能提供较多的工作岗位;

另一方面是因为尽管它们之间的物理距离接近,但是乘坐地铁需要换乘,导致实际的旅行距离增大. 本文认为早高峰时段乘客在这些区域间旅行,可能更倾向于选择公交甚至步行等更灵活的交通方式.

4.3 晚高峰时段区域功能分析

晚高峰客流峰段为 17:00—19:30,此时出行的客流大部分是市民下班产生的通勤客流,乘客的主要移动方向是由工作区去往居住区. 分析晚高峰时段的乘客在区域间的流动,结果如图 10 所示:

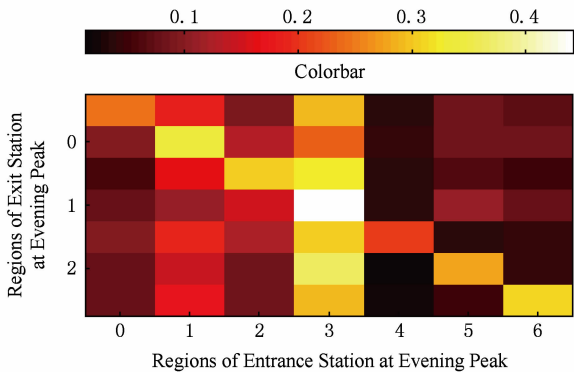


Fig. 10 Ridership statistics between regions at evening peak.

图 10 晚高峰乘客区域出行比例统计图

图 10 中矩形的每一行表示晚高峰该区域出站客流来自其他区域的比例. 发现每个区域的出站客流主要来自区域 3、本身区域和区域 1,表明在晚高峰阶段,区域 3 和区域 1 在释放了早高峰吸引的客流,客流回到自己的居住区.

4.4 晚低峰时段区域功能分析

晚低峰客流峰段为 19:30—23:30,此时出行的客流大部分是由夜间的娱乐休闲场所去往居住区. 本文统计了晚高峰和晚低峰各个区域的平均每天进站客流,如表 4 所示:

Table 4 Entrance Flow Statistics at Evening Peak and Evening Flat

表 4 晚高峰和晚低峰各个区域的进站客流

Region	Entrance Flow of Evening Peak per Day	Entrance Flow of Evening Flat per Day	Evening Peak/Evening Flat
3	221 819	183 082	0. 82
1	118 794	85 214	0. 72
2	96 721	60 312	0. 62
5	70 786	50 634	0. 72
0	65 509	51 239	0. 78
6	63 486	47 661	0. 75
4	27 888	22 130	0. 79

由晚高峰到晚低峰,区域 3 的进站客流下降的幅度小,不排除部分是因为在区域 3 乘客加班的因素影响,更可能的解释是,区域 3 此时扮演了晚上娱乐休闲核心区域的功能,注意到王府井、西单、东四十条都是北京夜间的重要娱乐场所.

在晚低峰客流峰段内,娱乐休闲区域不仅仅只有区域 3,其他区域也会出现自身区域内部的娱乐中心,吸引了本区域内较多的娱乐休闲的客流. 本文统计了平均每天晚低峰时段去往区域 6 的客流进站站点排名,如表 5 所示:

Table 5 Group 6 Flow Statistics of Evening Flat

表 5 晚低峰时段去往区域 6 的客流进站站点排名

Station	Flow Count Coming to Group 6 per Day
Wudaokou	3 562
Xizhimen	2 388
Dazhongsi	2 126
Olympic Park	2 078
Xidan	1 924
Dongzhimen	1 810
Xi'erqi	1 408

此时区域 6 内部的五道口、大钟寺成为区域 6 内部的娱乐中心. 值得注意的是晚高峰由五道口去往区域 1 的客流仅有 842 人,说明五道口是区域 6 自身独有的娱乐中心.

4.5 区域 1 和区域 3 的差别分析

在早高峰时,区域 3 和区域 1 都吸引了来自北京市周边区域的上班客流,但是需要注意的是,由区域 3 去往区域 1 的客流量远远低于由区域 1 去往区域 3 的客流量. 在下班时情况又相反,由区域 3 去往区域 1 的客流量远远高于由区域 1 去往区域 3 的客流量.

通过观察区域的地理分布,发现区域 3 明显站点分布密集,站点之间的距离小,乘客出站后换乘公交的需求和可能性小. 区域 1 的站点分布稀疏,站点之间的距离大,乘客出站后换乘公交的需求和可能性大.

由晚高峰到晚低峰,区域 3 的进站客流下降的幅度小于区域 1 下降的幅度,说明区域 1 在晚上提供的娱乐休闲功能要弱于区域 3.

4.6 区域功能总结

综上所述,区域 1 和区域 3 是北京市轨道交通系统的核心区域. 这 2 个区域在早高峰时吸收了其他区域的大量客流,又在晚高峰和晚低峰将客流释放回其他区域. 其中,区域 3 的核心地位相比区域 1 更高.

在早、晚高峰时,有相当大一部分的客流是在各自区域内部流动,说明这些乘客居住和工作在同一区域,也就是说乘客倾向于居住在离工作近的地方。

在晚低峰时,区域 3 的进站客流下降的幅度小,扮演了晚上娱乐休闲核心区域的功能。同时各个区

域内部也产生了它们自身独有的娱乐休闲核心。

4.7 功能区与行政区划分析

根据中国统计年鉴 2012^[21],表 6 统计了北京市各个行政区的常住人口和单位从业人员,表 7 将分类结果中的各个区域进行比较。

Table 6 Population Statistics of Beijing Districts

表 6 北京市行政区划人口统计

District	$10^{-4} \times$ Resident Population	$10^{-4} \times$ Working Population	Workers Density/Workers $\cdot$ km $^{-2}$
Dongcheng	91	75.4	18 006
Xicheng	124	112.7	22 299
Chaoyang	365.8	186.2	4 092
Fengtai	217	95.1	3 110
Shijingshan	63.4	23.5	2 782
Haidian	340.2	221.8	5 149
Fangshan	96.7	25.1	126
Tongzhou	125	38.3	422
Shuiyi	91.5	52.7	516
Changping	173.8	38.4	286
Daxing	142.9	39.1	378
Mentougou	29.4	11.2	77
Huairou	37.1	14.7	69
Pinggu	41.8	13.2	139
Miyun	47.1	15.9	71
Yanqing	31.9	7.7	39

Table 7 Statistics of Beijing Districts and Station Groups

表 7 北京市行政区对应区域站点数

District	Region0	Region1	Region2	Region3	Region4	Region5	Region6
Dongcheng	7	0	0	12	1	0	0
Xicheng	0	0	5	14	1	2	0
Chaoyang	7	16	1	1	7	8	9
Fengtai	4	17	16	1	0	0	0
Shijingshan	0	0	0	0	0	4	0
Haidian	0	9	9	3	4	4	7
Fangshan	0	0	9	0	0	0	0
Tongzhou	4	0	0	0	0	7	0
Shuiyi	0	0	0	0	0	0	7
Changping	3	0	0	0	0	0	12
Daxing	6	0	9	0	0	0	0
Mentougou	0	0	0	0	0	0	0
Huairou	0	0	0	0	0	0	0
Pinggu	0	0	0	0	0	0	0
Miyun	0	0	0	0	0	0	0
Yanqing	0	0	0	0	0	0	0

通过观察发现,北京市东城区和西城区的就业人口最密集,朝阳区、海淀区、通州区、大兴区和昌平区

区的居住人口较多。同时海淀区、朝阳区的就业人口最多,密度相对较大。

北京地铁区域3主要包含东城、西城2区的地铁站,而区域1主要经过海淀区和朝阳区,在一定程度上证明了区域3和区域1作为工作区的结论.

5 总 结

本文以地铁客流数据为基础提取了乘客出行模式和地铁站客流模式,然后建立了基于LDA主题模型的区域功能聚类模型,将具有相似出行特征的地铁站聚类在一起.聚类实验结果体现了区域的功能和以及在不同时段内区域的客流关系.

聚类结果中同一条线路上的地铁站倾向于聚类在一起,反映了同一线路上地铁站的客流相似性.而换乘站由于鉴于两种线路的客流特点,往往会成为区域转换的节点.下一步可以针对线路客流相似性和换乘站客流特性进行深入的挖掘分析.

参 考 文 献

[1] Antikainen J. The concept of functional urban area [J]. Findings of the Espo Project, 2005, 1(1): 447-456

[2] Wilkinson G G. Results and implications of a study of fifteen years of satellite image classification experiments [J]. IEEE Trans on Geoscience and Remote Sensing, 2005, 43(3): 433-440

[3] Herold M, Liu X, Clarke K C. Spatial metrics and image texture for mapping urban land use [J]. Photogrammetric Engineering and Remote Sensing, 2003, 69(9): 991-1002

[4] Liu L, Biderman A, Ratti C. Urban mobility landscape: Real time monitoring of urban mobility patterns [C] //Proc of the 11th Int Conf on Computers in Urban Planning and Urban Management. Piscataway, NJ: IEEE, 2009: 1-16

[5] Goh C Y, Dauwels J, Mitrovic N, et al. Online map-matching based on Hidden Markov model for real-time traffic sensing applications [C] //Proc of the 15th IEEE Conf on Intelligent Transportation Systems. Piscataway, NJ: IEEE, 2012: 776-781

[6] Liu W, Zheng Y, Chawla S, et al. Discovering spatio-temporal causal interactions in traffic data streams [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1010-1018

[7] Guc B, May M, Saygin Y, et al. Semantic annotation of GPS trajectories [C] //Proc of 11th AGILE Int Conf on Geographic Information Science. Berlin: Springer, 2008: 1-9

[8] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 186-194

[9] Pan G, Qi G, Wu Z, et al. Land-use classification using taxi GPS traces [J]. IEEE Trans on Intelligent Transportation Systems, 2012, 14(1): 113-123

[10] Li X, Pan G, Wu Z, 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

[11] Sheng C, Zheng Y, Hsu W, et al. Answering top-*k* similar region queries [C] //Proc of the 15th Database Systems for Advanced Applications. Berlin: Springer, 2010: 186-201

[12] Xue A Y, Zhang R, Zheng Y, et al. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction [C] //Proc of the 29th IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2013: 254-265

[13] Zheng Y, Xie X. Learning location correlation from gps trajectories [C] //Proc of the 29th IEEE Int Conf on Mobile Data Management. Piscataway, NJ: IEEE, 2010: 27-32

[14] Zheng Y, Zhang L, Xie X, et al. Mining correlation between locations using human location history [C] //Proc of the 17th ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2009: 472-475

[15] Chen C, Chen J, Barry J. Diurnal pattern of transit ridership: A case study of the New York City subway system [J]. Journal of Transport Geography, 2009, 17(3): 176-186

[16] Leng B, Zeng J, Xiong Z, et al. Probability tree based passenger flow prediction and its application to the Beijing subway system [J]. Frontiers of Computer Science, 2013, 7(2): 195-203

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

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

[19] Kruschke J K. Bayesian data analysis [J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2010, 1(5): 658-676

[20] Azzopardi L, Girolami M, van Risjbergen K. Investigating the relationship between language model perplexity and IR precision-recall measures [C] //Proc of the 26th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2003: 369-370

[21] National Bureau of Statistics of China. China Statistics Yearbook 2012 [M]. Beijing: China Statistics Press, 2013



Leng Biao, born in 1982. Associate professor. Member of China Computer Federation. His research interests include intelligent transportation system, smart city, data vitalization, pattern recognition, computer vision.



Zhao Wenyuan, born in 1990. Master. His research interests include intelligent transportation system.