Journal of Computer Applications
计算机应用,2017,37(8):2362 - 2367
ISSN 1001-9081
CODEN JYIIDU
2017-08-10
http://www.joca.cn

文章编号:1001-9081(2017)08-2362-06
doi:10.11772/j.issn.1001-9081.2017.08.2362

基于多源出行数据的居民行为模式分析方法

徐晓伟[1,2],杜　一[1*],周园春[1]
(1.中国科学院计算机网络信息中心 大数据技术与应用发展部,北京 100190;　2.中国科学院大学,北京 100049)
(*通信作者电子邮箱 duyi@cnic.cn)

摘　要:基于对智能交通卡数据的挖掘与分析能够为城市交通建设和城市管理提供有力支持,但现有研究数据大都仅包含公交或地铁这两方面数据,且主要关注群体性宏观出行规律。针对这一问题,以某城市交通卡数据为例,该数据包含着城市居民日常出行公交、地铁、出租车等多源数据,首先提出行程链的概念对居民出行行为建模,在此基础上给出不同维度的周期性出行特征;然后提出一种基于最长公共子序列的空间周期性特征提取方法,并对城市居民出行规律进行聚类分析;最后通过规则定义5个评价指标对该方法的有效性进行初步验证。结果表明引入该方法的聚类算法对聚类结果有6.8%的效果提升,有利于发现居民的行为模式。

关键词:智能交通卡;多源数据;序列匹配;聚类分析;时空数据挖掘
中图分类号:TP391.4;TP181　　**文献标志码**:A

Resident behavior model analysis method based on multi-source travel data

XU Xiaowei[1,2], DU Yi[1*], ZHOU Yuanchun[1]
(1. *Department of Big Data Technology and Application Development, Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: The mining and analysis of smart traffic card data can provide strong support for urban traffic construction and urban management. However, most of the existing research data only include data about bus or subway, and mainly focus on macro-travel patterns. In view of this problem, taking a city traffic card data as the example, which contains the multi-source daily travel data of urban residents including bus, subway and taxi, the concept of tour chain was put forward to model the behavior of residents. On this basis, the periodic travel characteristics of different dimensions were given. Then a spatial periodic feature extraction method based on the longest common subsequence was proposed, and the travel rules of urban residents were analyzed by clustering analysis. Finally, the effectiveness of this method was verified by five evaluation indexes defined by the rules, and the clustering result was improved by 6.8% by applying the spatial periodic feature extraction method, which is helpful to discover the behavior pattern of residents.

Key words: smart traffic card; multi-source data; sequence matching; clustering analysis; spatio-temporal data mining

0　引言

当前中国城市化进程发展迅速,由于人口数量的日益增长造成城市交通拥堵、空气污染等问题,给城市居民的工作生活带来了严重的影响。研究城市公共交通日益成为解决城市各种交通问题的主要方法[1],而交通卡数据也逐渐受到越来越多学者的关注。交通卡数据属于基于事件触发的轨迹数据,即移动对象触发传感器事件后而被记录下来形成的轨迹[2]。乘客每次上下车或进出站刷卡都会触发传感器而记录当前的数据信息,在一些城市使用交通卡可以乘坐地铁、公交、出租车等,这些数据隐藏着大量有价值的信息,如城市居民日常工作生活的出行模式等。与公共交通数据相关的交通部门、旅游公司以及研究学者可以通过利用该数据对居民的出行模式进行建模与分析,从而挖掘出有价值的信息。

基于交通卡数据的研究主要集中在:出行记录的起讫点矩阵(Origin, Destination Matrix, OD Matrix)推断、居民出行特征研究、职业住宅区域识别和异常模式中特定人群发现等方面。在出行记录的起讫点矩阵分析与推断方面,Munizaga等[3]结合交通卡数据和全球定位系统(Global Positioning System, GPS)数据,提出一种基于公交站点级别的OD矩阵推断下车点;Abrahamsson[4]通过分析交通网络中不同站点之间连接权重,即交通流量,计算OD矩阵;Wong等[5]提出一种基于信息熵的方法基于观察到的客流量分析有时间依赖的OD矩阵。在居民行为特征研究方面,Ma等[6]研究北京市交通卡数据(公交和地铁),对空间数据进行DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类后,基于一些时空特征分析乘客的出行规律;Liu等[7]分别从时间和空间维度上研究深圳市城市居民的宏观出行规律等;李海波等[8]基于苏州市的交通卡数据的时间特征对三大人群进行研究。Bagchi等[9-10]基于交通卡数据分析研究了城市居

收稿日期:2017-02-13;**修回日期**:2017-04-27。
基金项目:国家重点研发计划项目(2016YFB0501900, 2016YFB1000600);国家自然科学基金资助项目(61402435)。
作者简介:徐晓伟(1993—),男,河北邯郸人,硕士研究生,主要研究方向:数据挖掘、机器学习;　杜一(1988—),男,山东聊城人,副研究员,博士,CCF会员,主要研究方向:数据挖掘、数据可视化;　周园春(1975—),男,江西鹰潭人,研究员,博士,CCF会员,主要研究方向:大数据管理与处理技术、数据挖掘。

民每张交通卡的中转率、出行率以及使用智能交通卡对链式出行的影响。在识别职业住宅区域研究上,文献[11-12]基于规则和决策树方法分析城市的公交或地铁 OD 数据模型,用于发现该城市的职业住宅区域。将交通卡数据应用到识别特定人群方面,Xue 等[13]分析了新加坡地铁数据,将地铁各个站点对游客的吸引度作为先验知识,人工标识一些游客数据,使用强化迭代算法发现通行者中的游客,促进旅游业的发展;Du 等[14]结合交通卡数据、地理空间数据(路网信息)和兴趣点(Point of Interest, POI)数据等,提取大量特征用于发现公共出行中的小偷。

目前国内外学者对交通卡数据的研究主要集中在宏观群体上,同时个体行为特征研究的文章局限于公交车或地铁数据,对于多源数据的融合分析研究不够。针对上述数据的多源性以及个体出行特征研究不足等问题,本文基于多种数据源研究城市居民的个体出行模式,将最长公共子序列(Longest Common Subsequence, LCS)的方法[15]应用于乘客每天的出行轨迹,充分提取出数据中的空间信息,综合出行数据的时空特征分析居民的出行模式。本文提出一种有效的轨迹序列匹配方法,不再关注固定的参考轨迹序列,将乘客在这段时间的所有轨迹序列进行两两匹配,计算出所有轨迹序列之间的最长公共子序列的平均相似度情况。

1　基于出行场景的数据清洗

1.1　交通卡数据结构

本文研究基于某城市 2015 年 4 月中某 8 天的交通卡数据。在该城市,交通卡可以用于乘坐公交、地铁、出租、轮渡和 P+R停车场。在各种出行方式中,地铁占 59.9%,公交占 37.7%。其中:地铁乘车记录具有完整的进出站信息;公交记录了上车时间、线路和消费金额。每一条出行记录包括交通卡编号、乘车日期、乘车时间、乘车站点、乘车类型、消费金额和是否优惠。本文假设一张交通卡对应一名乘客,即乘客编号。数据中各字段及对应含义如下:

1)交通卡编号(id):用于标识交通卡(乘客)的唯一编号;

2)乘车日期(date):乘客刷卡上下车日期,格式为 YYYY-MM-DD;

3)乘车时间(time):乘客刷卡上下车时间,为 24 小时制;

4)乘车站点(station):若类型为公交或地铁,则为公交线路或地铁站点,其他为无;

5)乘车类型(type):公交、地铁、出租、轮渡、P+R 停车场等多种乘车方式;

6)消费金额(price):此次记录消费的金额;

7)是否优惠(discount):根据交通部门制定的一些优惠政策对消费记录进行判定。

1.2　交通卡数据清洗

城市交通卡数据由于网络传输、设备故障等原因存在数据丢失或产生脏数据,本文给出了对该类数据的清洗流程,如图 1 所示。数据清洗流程中给出了对应的清洗规则,该规则结合了该城市数据采集机制、刷卡优惠政策等,对数据的合理性进行评估。首先对交通卡数据按照交通卡编号和乘车时间排序;然后检查数据中是否存在时间格式错误的记录以及针对每名乘客每天是否存在单条地铁数据,若发现将其剔除;最后经过优惠政策的筛选得到清洗后的数据。

图 1 中的优惠政策过滤模块主要针对地铁乘车记录。乘坐地铁每刷一次卡就会生成一条记录,乘坐地铁进出站各刷卡一次为完整的记录,因此处理数据时按照每两条进行处理,获取所有乘客列表,遍历每名乘客的乘车记录,具体流程如图 2,其中 cnt 表示该用户的总记录数,idx 表示当前乘客第 idx 条记录,$idx+1<cnt$ 表示当前处理的下一条记录序号小于记录总数。由于判断条件较多,为方便表示,图 2 中判断缺少的处理默认为:跳过当前记录,读取下一行。

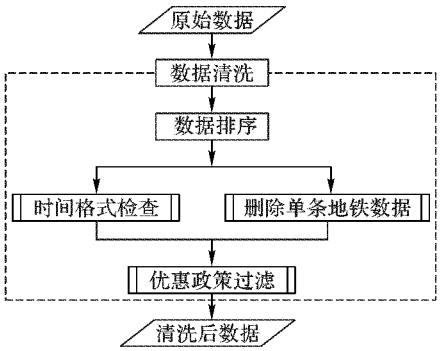

图 1　交通卡数据清洗流程

Fig. 1　Cleaning process of traffic card data

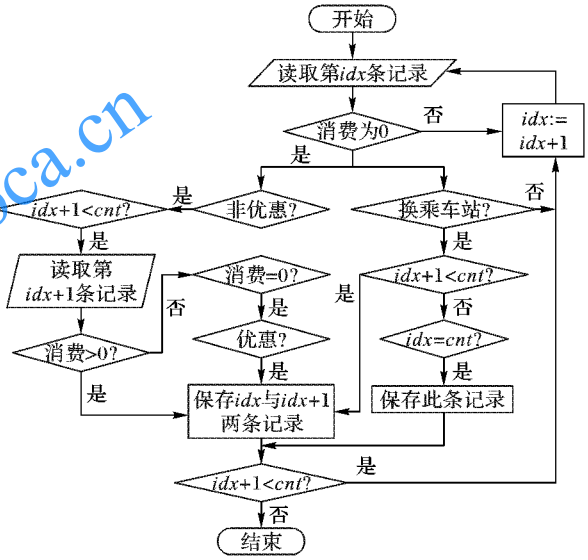

图 2　优惠政策处理流程

Fig. 2　Process of preferential policy

图 2 中出站换乘规则具体如下:若在换乘站点出站后在间隔 30 min 内再进站可以享受出站换乘连续计费。例如,某名乘客的一日乘坐地铁记录,共计 n 条,其中第 i 条记录为:

$$R_i:\{uid,time_i,station_i\};\ 1\leqslant i\leqslant n$$

由于地铁记录进出站点刷卡两次为一条完整出行记录,所以按照每两条数据进行处理,一般情况下从 $station_0$ 站出发,中间经过站点 $station_{i-1}\in$“换乘车站”出站,并且在 30 min内再次从该站点进站(即 $time_i-time_{i-1}\leqslant 30$ min),最后到达 $station_n$ 站。系统会将票价连续计费,计算从 $station_0$ 站到 $station_n$ 站的总费用。若 n 为奇数,且最后两条记录乘车站点为换乘车站,则说明最后一条表示从换乘车站出站 。

2　基于 LCS 匹配的出行模式特征提取

基于给定的清洗后的数据,本章将给出基于最长公共子序列(LCS)匹配的出行模式特征提取。首先对轨迹与站点序列等概念进行定义,然后给出基于站点序列等相似性度量方

法。相似性度量表示乘客在一定时间里周期性空间特征规律程度。

2.1 基本定义

定义 1 轨迹。表示在一段时间内按照时间排序乘客出行的位置集合，$L = \{(l_1,t_1),(l_2,t_2),\cdots,(l_i,t_i),\cdots,(l_n,t_n)\}$。其中：$l_i$ 表示第 i 个位置；t_i 表示在 i 个位置时发生的时间，$t_{i-1} \leqslant t_i$；n 表示在该条轨迹中经过的位置个数，$1 \leqslant i \leqslant n$。

定义 2 站点序列。将轨迹中的数据元素抽取出时间节点后留下的有序集合，表示为 $L = \{l_1,l_2,\cdots,l_i,\cdots,l_n\}$，在本文中位置 l_i 表示地铁站点、公交线路等。

2.2 路线序列的相似性度量

基于前文的定义 1 和 2，本文抽取出每名乘客在这段时间里每天出行轨迹中的站点序列，提出一种基于最长公共子序列匹配的出行模式特征提取方法，对乘客每日的乘车站点序列进行相似性度量，从而为判断每名乘客的出行模式是否有规律提供依据。

假设 $A = \{l_{a1},l_{a2},\cdots,l_{ai},\cdots,l_{an}\}$ 与 $B = \{l_{b1},l_{b2},\cdots,l_{bi},\cdots,l_{bm}\}$ 表示在这段时间内任意两天中的站点序列，其中 $n(1 \leqslant i \leqslant n)$ 表示站点序列 A 的长度，$m(1 \leqslant j \leqslant m)$ 表示站点序列 B 的长度，使用 $D[i,j]$ 表示路线序列 A 与 B 之间最长公共子序列的长度，即：

$$D[i,j] = \begin{cases} 0, & i = 0 \text{ 或 } j = 0 \\ D[i-1,j-1]+1, & i > 0, j > 0, l_{ai} = l_{bj} \\ \max\{D[i,j-1],D[i-1,j]\}, & i > 0, j > 0, l_{ai} \neq l_{bj} \end{cases} \tag{1}$$

求得任意两天中的站点序列间最长公共子序列长度后，进行相似性度量 $LSim(A,B)$，使用公共子序列出现的频率作为度量方式，若值越大，则代表站点序列之间经过的相同地方越多，相似性越高。假设出行天数为 N，在这 N 天里该乘客的出行轨迹路线的相似性评分为 $score$，其计算公式如下：

$$LSim(A,B) = D[n,m]/(n+m) \tag{2}$$

$$score = \frac{1}{N}\sum_{A}\sum_{B \neq A} LSim(A,B) \tag{3}$$

下面给出该方法的具体算法：首先将交通卡数据按照乘客编号分类汇总，抽取出这段时间的出行站点序列；然后遍历并处理每名乘客的出行记录，若该乘客的出行记录总数 $length = 1$，由于无法比对，赋值相似性评分 $score$ 为 0，其他情况下对每日出行站点序列两两匹配计算基于最长公共子序列的分数，求取这些分数的平均值作为该用户的最终结果。

算法 1 基于最长公共子序列出行站点序列相似性计算。

输入 乘客出行记录数据数组 $Records$，长度为 num，每个元素代表对应乘客的记录；

输出 乘客出行站点序列的相似性评分数组。

```
for idx := 0 to num - 1
    user := Records[idx]
    length := user.length
    score[idx] = 0
    if total := 1
        continue
    end if
    for i := 0 to length
        for j := i + 1 to length
            score := score + LCS(user[i], user[j])
        end for
    end for
    score[idx] := score[idx]/length
    return score
end for
```

3 出行数据聚类

本章首先给出出行记录(Time-Station-Duration，TSD)模型和行程链(Tour)模型，用来表征数据中的关键信息；然后从时间、空间、行程链等维度分析周期性提取特征；最后使用 *K*-Means ++ 算法[16]进行聚类计算。

3.1 数据模型

本文构建了出行记录(TSD)和行程链(Tour)两个数据模型。出行记录表示乘客的一次上下乘车过程，由于数据多源性，包含了公交、出租、地铁等类型，具体定义如下：$TSD = \{T,S,D\}$。其中：T 表示乘客在一次出行记录中的上车时间；S 表示出行记录中的站点信息，若为地铁出行，站点包括上下车站点，若为公交出行，站点为公交线路，其他为对应的出行方式；D 表示此次记录乘车时间。

行程链这一数据模型用于表达预处理后的数据，定义为时间相近的出行记录集合，某个行程链 $Tr = \{tsd_1.T,tsd_n.T + tsd_n.D,(tsd_i.S)\}$，其中 tsd_i 表示该名乘客的第 $i(1 \leqslant i \leqslant n)$ 条出行记录。考虑到乘客从某个起点到某个终点需要经过多条出行记录，若上一条出行记录的下车时间与下一条出行记录上车时间之间的时间间隔小于某一阈值 ε，即 $tsd_{i+1}.T - tsd_i.T \leqslant \varepsilon$，则将这两条记录进行合并，一直循环操作，直到出行记录之间的时间间隔超出阈值，整理生成该乘客在该天的行程。

由于数据具有多源性，对于除地铁数据以外的其他类型数据均无乘车过程的详细时间，为解决这一问题，对于每条无乘车持续时间的记录统一设置为 30 min，同时在本文数据处理过程中给定相邻的出行记录之间是否同属于同一行程的阈值为 $\varepsilon \leqslant 30$ min。

3.2 特征提取

为了更好地挖掘出交通卡数据中个体用户的出行规律模式，本文选取并生成了一些特征属性，分别从时间、空间、行程链等方面对每名乘客出行模式的周期性进行有效的量化计算，各个特征解释如下：*startTime* 表示乘客每天第一次上车时间；*endTime* 表示每天最后下车时间；*tourSD* 表示基于行程链的概念计算行程次数方差；*total* 表示在这段时间里的出行次数；*stopSeq* 表示这段时间里出行站点的相似性比较。

3.2.1 时间周期性特征计算方法

用列表存储每个乘客早晚的出行时间，按照众数思想计算出每个乘客的出行时间参照点，然后将时间列表与该参照点进行比较分析。本文假设规律出行的乘客每日出行时间中的小时时刻值是固定的，且时间波动前后不超过 30 min，因此以出行时间中小时值的众数作为参考值。以 *startTime* 为例，计算流程如图 3，若出行时间中小时的众数个数等于 1，即表示该乘客没有规律的出行时间，则置 *startTime* 为 1/*length*(*length* 为出行总次数)。

对于小时时刻值的众数个数大于 1 的计算方法如下：

1)众数选择基于小时时刻值，相同的小时作为参照时间中的小时值；

2)然后计算这些相同小时的时间点中分钟的平均数，从而得到参照时间点；

3)遍历列表中的时间与参照时间点，设置阈值为 30 min，

若时间间隔小于等于 30 min,则计数 *startCnt* + = 1,计算百分比 *startCnt/length*(*startCnt* 为规律出行的次数)。

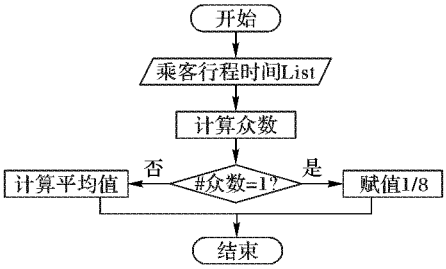

图 3　出行时间周期性计算流程

Fig. 3　Calculation process of the periodicity of travel time

3.2.2　*tourSD* 计算方法

图 4 为乘客行程链次数的周期性计算流程:首先筛选出所有出行次数大于 1 的乘客,计算出它们行程链次数的方差;然后根据整体数据分布情况对出行次数为 1 的乘客进行数据填充,计算过程中发现该特征取值范围为 0 到 5.4,平均值为 0.43,平均值对于填充出行一次的乘客不准确,在本文中选取数据中的 95% 位,取值 1.2。

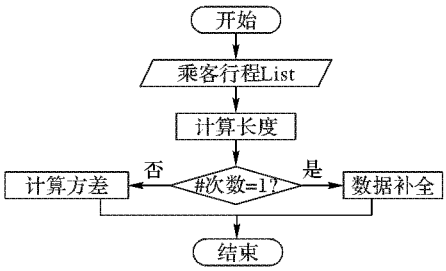

图 4　行程链周期性计算流程

Fig. 4　Calculation process of the periodicity of tour chain

由于各个特征指标的取值范围不同,本文使用归一化将不同特征的数值约束到[0,1]区间,归一化方法公式如下:

$$x = (x - x_{\min})/(x_{\max} - x_{\min})$$

其中特征属性 *startTime*、*endTime*、*total* 和 *stopSeq* 表征着乘客在这段时间里出行模式的相似程度,值越大代表越规律;而 *tourSD* 表示乘客出行行程链的方差,值越小表示出行越规律,为保持所有特征值与规律的正向关联性,经过归一化处理后,对于特征 *tourSD* 作简单处理,$x_{tourSD} := 1 - x_{tourSD}$。

3.3　规律性出行聚类

K-Means 算法是经典的聚类方法,但存在以下不足:1)聚类结果容易受初始值选取的影响,不同的随机种子点会得到完全不同的结果;2)算法复杂度较高,运算时间较长。为应对以上问题,本文使用 *K*-Means 的一种改良算法——*K*-Means ++。该算法有效地解决了 *K*-Means 初始化随机种子的不足,它初始化随机种子的原则为选取的种子点之间距离越大,被选取作为聚类中心的概率越大,这样有利于获取数据分布信息,实现更好的结果。

3.3.1　*k* 值选择

本文将空间周期性特征 *stopSeq* 作为控制变量,分别运行两次聚类算法进行比对,其中 cluster0 采用前 4 个特征,cluster1 采用全部 5 个特征(添加了 *stopSeq*),对于 *k* 值(聚类个数)的选取根据类内距离误差和,设定 *k* 值范围为 2 到 9,针对每个 *k* 值多次计算求取平均值,然后根据每个类的类内距离误差变化情况求得合适的 *k* 值。

图 5(a)、(b)分别为 cluster0 与 cluster1 的 *k* 与类内距离误差和之间的分布图,可以发现当 *k* = 5 时,误差和变化出现较大的拐点,因此将 *k* 赋值为 5,对于五类按照出行规律分别表示为:非常规律(VH)、一般规律(H)、中等(M)、一般不规律(L)和不规律(VL)。

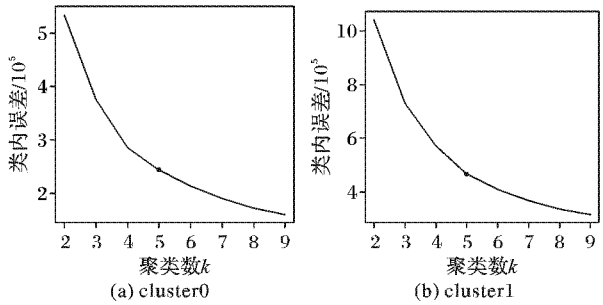

图 5　聚类个数 *k* 与类内距离误差和分布图

Fig. 5　Distribution of number of clusters (*k*) and sum square error within clusters

3.3.2　聚类结果展示

表 1 和表 2 分别为 cluster0 与 cluster1 的聚类中心分布表,图 6 展现了 cluster0 与 cluster1 的聚类结果。由图 6 中可以看出,对应表中数值越大代表出行模式越规律,根据不同的规律程度,该算法将城市居民划分成 5 个非常明显的层次,且发现从中等(M)到一般不规律(L)数值下降非常明显。同时也发现引入站点序列周期性相似特征的 cluster1 中前两类各项特征明显高于 cluster0,这说明通过引入站点序列相似性度量这一特征使得聚类结果中的中心点更加准确。

表 1　cluster0 聚类中心分布

Tab. 1　Clustering center distribution of cluster0

特征	出行规律				
	VH	H	M	L	VL
startTime	0.565	0.346	0.151	0.025	0.001
endTime	0.449	0.211	0.120	0.021	0.001
tourSD	0.728	0.708	0.426	0.188	0.009
total	0.917	0.869	0.882	0.938	0.757

表 2　cluster1 聚类中心分布

Tab. 2　Clustering center distribution of cluster1

特征	出行规律				
	VH	H	M	L	VL
startTime	0.536	0.361	0.092	0.064	0.003
endTime	0.409	0.247	0.069	0.054	0.003
tourSD	0.730	0.672	0.318	0.234	0.032
total	0.930	0.861	0.890	0.888	0.793
stopSeq	0.718	0.696	0.339	0.231	0.001

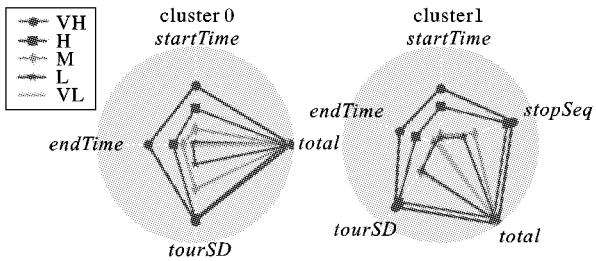

图 6　聚类中心雷达图

Fig. 6　Radar map of clustering center

表 3 展示了两种聚类方法关于不同规律程度的人群分布情况,其中 cluster0 聚类结果中前两类(出行非常规律和一般规律)约占 26%,cluster1 中前两类约占 28%,说明通过引入站点序列相似性度量这一特征进一步挖掘出了聚类结果中更

多规律出行的人群。

表 3 两种聚类方法与不同出行规律的人群分布

Tab. 3 Crowd distribution with different travel regularity for two clustering methods

出行规律	cluster0	cluster1	出行规律	cluster0	cluster1
VH	1 218 398	1 183 249	L	2 448 223	2 107 980
H	1 234 603	1 453 954	VL	3 021 509	3 618 523
M	1 503 118	1 062 145			

4 出行数据聚类结果评价

为了更好地说明聚类结果的有效性，本文通过选取一些重要的周期性指标对城市居民进行规律聚类评估，如对于规律出行的人群常常基于“家—公司—家”或者“家—学校—家”模式，分析在出行记录间是否有长时间间隔作为指标等。为方便解释，本文基于前面提到的数据模型，假设一共有 N 名乘客，每名乘客出行 n 天，第 i 日的行程链为 Tr_i，每日出行记录为 m 条。

$$\begin{cases} Tour = \{Tr_1, Tr_2, \cdots, Tr_i, \cdots, Tr_n\};\ 1 \leqslant i \leqslant n \\ Tr_i = \{tsd_{i1}.T, tsd_{im}.T + tsd_{im}.D, (tsd_{ij}.S)\};\ 1 \leqslant j \leqslant m \end{cases}$$

指标的具体介绍如下：

1) *CalTimeOD*：规律出行的人群多见于上班族、学生等群体，他们出行轨迹的规律常常是 ABA 模式，即“家—公司—家”或者“家—学校—家”，在锚点 B 上停留的时间超过 6 h。因此该指标要求时间上存在相邻的出行记录且出行时间间隔超过 6 h，并且空间上要求出行轨迹的首尾站点相同，计算在这段时间里符合规则的百分比，其中 $rule_i$ 表示乘客第 i 日对应该指标的出行情况。

$$\begin{cases} rule_i: tsd_{i1}.S = tsd_{im}.S \text{ 且 } tsd_{ij+1}.T - tsd_{ij}.T \geqslant 6\text{ h} \\ calTimeOD = \dfrac{1}{N}\sum_N \sum_n \dfrac{1\{rule_i\}}{n} \end{cases} \tag{4}$$

2) *CalTimeODLabel*：原理同上，在评价每名乘客时若该乘客在自己的出行记录中有一半时间以上满足上述规则，则定义为出行规律的人，从而计算聚类后结果的准确率。

$$\begin{cases} rule_i: tsd_{i1}.S = tsd_{im}.S \text{ 且 } tsd_{ij+1}.T - tsd_{ij}.T \geqslant 6\text{ h} \\ calTimeODLabel = \dfrac{1}{N}\sum_N 1\left\{\sum_n \dfrac{1\{rule_i\}}{n} \geqslant \dfrac{1}{2}\right\} \end{cases} \tag{5}$$

3) *CalOD*：该指标分析每名乘客所有的出行轨迹中频率最高的轨迹出发点 $freq.S_{start}$ 与结束点 $freq.S_{end}$，作为参考轨迹的开始点与结束点；然后将该乘客的每条轨迹与参考轨迹的起始点进行比较，计算在这段时间里符合规则的百分比。

$$\begin{cases} rule_i: tsd_{i1}.S = freq.S_{start} \text{ 且 } tsd_{im}.S = freq.S_{end} \\ calOD = \dfrac{1}{N}\sum_N \sum_n \dfrac{1\{rule_i\}}{n} \end{cases} \tag{6}$$

4) *CalPalindScore*：有规律的乘客出行轨迹应该为标准的回文序列，依据每条轨迹中的回文程度给予对应的分数，使用 $palindrome(Tr_i)$ 表示第 i 天出行轨迹周期性分数，从而计算出该乘客在这段时间里出行规律的最终分数。

$$calPalindScore = \frac{1}{N}\sum_N \sum_n \frac{palindrome(Tr_i)}{n} \tag{7}$$

5) *CalPalindLabel*：原理同上，不同点在于若该乘客在这段时间里的评分分数超过 0.5，则认为乘客规律出行，从而计算聚类后结果的准确率。

$$calPalindLabel = \frac{1}{N}\sum_N 1\left\{\sum_n \frac{palindrome(Tr_i)}{n} \geqslant \frac{1}{2}\right\} \tag{8}$$

表 4 针对上述提到的 5 个指标分别进行了测试，每一种指标有三组测试数据。测试数据根据 5 类聚类结果规律程度组合划分为：非常规律(VH)，非常规律(VH) + 一般规律(H)，非常规律(VH) + 一般规律(H) + 中度(M)。

分析表 4 发现：1)逐步约束测试数据规律程度的范围，两种聚类方法都逐渐增大规律人群所占的比例；2)定义 CalTimeODLabel、CalPalindLabel 指标的结果高于同样规则下计算具体分数值指标，是因为计算具体分数指标的方法对于评价乘客是否出行规律有较大的约束；3)引入基于 LCS 的站点序列周期性特征后的聚类结果整体上都明显好于前一个聚类结果，从 VH 规律程度上分析准确率平均提高 6.8 个百分点，这充分说明引入该特征后更有利于准确地发现居民中规律人群。

表 4 两种聚类方法基于不同指标评价结果

Tab. 4 Evaluation results of two clustering methods based on different indexes

评价指标	规律组合	Cluster0	Cluster1
calTimeOD	VH + H + M	0.192	0.212
	VH + H	0.294	0.246
	VH	0.474	0.534
calTimeODLabel	VH + H + M	0.175	0.202
	VH + H	0.267	0.261
	VH	0.522	0.607
CalOD	VH + H + M	0.496	0.535
	VH + H	0.540	0.572
	VH	0.637	0.729
CalPalindScore	VH + H + M	0.573	0.585
	VH + H	0.590	0.587
	VH	0.640	0.681
CalPalindLabel	VH + H + M	0.621	0.635
	VH + H	0.638	0.642
	VH	0.708	0.768

5 结语

本文基于交通卡多源数据引入一种根据最长公共子序列的计算空间周期性特征方法，使得特征数据包含更加丰富的时空信息，在聚类分析中将居民依据不同的规律程度划分为 k 个等级。实验中根据交通卡数据特征给出了 5 个聚类指标评价，发现结合该特征后的聚类结果效果更好。研究城市居民的出行特征能帮助我们发现居民的出行模式，洞察出数据中隐藏的信息，在下一步工作中，将根据公共交通出行数据提取出更多详细的特征，进一步提升聚类效果。除此之外，将聚类结果应用到具体场景中，如依据规律出行的人群发现职业住宅区域，根据异常出行模式识别特殊人群等，也将作为下一步工作。

参考文献 (References)

[1] 龙瀛，孙立君，陶遂. 基于公共交通智能卡数据的城市研究综述[J]. 城市规划学刊，2015(3)：70－77. (LONG Y, SUN L J, TAO S. A review of urban studies based on transit smart card data [J]. Urban Planning Forum, 2015(3): 70－77.)

[2] 李婷，裴韬，袁烨城，等. 人类活动轨迹的分类，模式和应用研究综述[J]. 地理科学进展，2014，33(7)：938－948. (LI T, PEI T, YUAN Y C, et al. A review on the classification, patterns and applied research of human mobility trajectory [J]. Progress in Geography, 2014, 33(7): 938－948.)

[3] MUNIZAGA M A, PALMA C. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile [J]. Transportation Research Part C:

Emerging Technologies, 2012, 24: 9 – 18.
[4] ABRAHAMSSON T. Estimation of origin-destination matrices using traffic counts — a literature survey, IR-98-021 [R]. Laxenburg, Austria: International Institute for Applied Systems Analysis, 1998.
[5] WONG S C, TONG C O. Estimation of time-dependent origin-destination matrices for transit networks [J]. Transportation Research Part B: Methodological, 1998, 32(1): 35 – 48.
[6] MA X, WU Y-J, WANG Y, et al. Mining smart card data for transit riders' travel patterns [J]. Transportation Research Part C: Emerging Technologies, 2013, 36: 1 – 12.
[7] LIU L, HOU A, BIDERMAN A, et al. Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen [C]// ITSC '09: Proceedings of the 2009 12th International IEEE Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE, 2009: 1 – 6.
[8] 李海波, 陈学武, 陈峥嵘. 基于公交 IC 卡数据的乘客出行时间特征研究[C]//中国城市交通规划年会暨第 27 次学术研讨会. 北京: 中国城市规划学会, 2014. (LI H B, CHEN X W, CHEN Z R. Study on passenger travel time characteristics based on public traffic IC card data [C]// Proceedings of the 2014 China Urban Transport Planning Annual Conference & 27th Symposium. Beijing: Urban Planning Society of China, 2014.)
[9] BAGCHI M, WHITE P. What role for smart-card data from bus systems? [J]. Municipal Engineer, 2004, 157(1): 39 – 46.
[10] BAGCHI M, WHITE P R. The potential of public transport smart card data [J]. Transport Policy, 2005, 12(5): 464 – 474.
[11] 龙瀛, 张宇, 崔承印. 利用公交刷卡数据分析北京职住关系和通勤出行[J]. 地理学报, 2012, 67(10): 1339 – 1352. (LONG Y, ZHANG Y, CUI C Y. Identifying commuting pattern of Beijing using bus smart card data [J]. Acta Geographica Sinica, 2012, 67(10): 1339 – 1352.)
[12] 许志榕. 上海市职住关系和通勤特征分析研究——基于轨道交通客流数据视角[J]. 上海城市规划, 2016(2): 114 – 121. (XU Z R. Study on job-housing relationship and characteristic of commuting in Shanghai: based on the perspective of rail transit passenger flow data [J]. Shanghai Urban Planning Review, 2016(2): 114 – 121.)
[13] XUE M, WU H, CHEN W, et al. Identifying tourists from public transport commuters [C]// KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1779 – 1788.
[14] DU B, LIU C, ZHOU W, et al. Catch me if you can: detecting pickpocket suspects from large-scale transit records [C]// KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 87 – 96.
[15] VLACHOS M, GUNOPULOS D, KOLLIOS G. Discovering similar multidimensional trajectories [C]// ICDE '02: Proceedings of the 18th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2002: 673 – 684.
[16] ARTHUR D, VASSILVITSKII S. *K*-means++: the advantages of careful seeding [C]// SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007: 1027 – 1035.

This work is supported by the National Key Research and Development Program (2016YFB0501900, 2016YFB1000600), the National Natural Science Foundation of China (61402435).

XU Xiaowei, born in 1993, M. S. candidate. His research interests include data mining, machine learning.

DU Yi, born in 1988, Ph. D., associate professor. His research interests include data mining, data visualization.

ZHOU Yuanchun, born in 1975, Ph. D., professor. His research interests include big data management and processing, data mining.

(上接第 2361 页)

[3] CAMPAGNA A, PAGH R. Finding associations and computing similarity via biased pair sampling [C]// ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2009: 61 – 70.
[4] ATTEYA W A, DAHAL K, HOSSAIN M A. Distributed BitTable multi-Agent association rules mining algorithm [C]// KES '11: Proceedins of the 15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, LNCS 6881. Berlin: Springer-Verlag, 2011: 151 – 160.
[5] 赵学健, 孙知信, 袁源, 等. 一种正交链表存储的改进 Apriori 算法[J]. 小型微型计算机系统, 2016, 37(10): 2291 – 2295. (ZHAO X J, SUN Z X, YUAN Y, et al. An improved apriori algorithm based on orthogonal storage [J]. Journal of Chinese Computer Systems, 2016, 37(10): 2291 – 2295.)
[6] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation [J]. ACM SIGMOD Record, 2000, 29(2): 1 – 12.
[7] 李也白, 唐辉, 张淳, 等. 基于改进的 FP-tree 的频繁模式挖掘算法[J]. 计算机应用, 2011, 31(1): 101 – 103. (LI Y B, TANG H, ZHANG C, et al. Frequent pattern mining algorithm based on improved FP-tree [J]. Journal of Computer Applications, 2011, 31(1): 101 – 103.)
[8] MA Z, YANG J, ZHANG T, et al. An improved eclat algorithm for mining association rules based on increased search strategy [J]. International Journal of Database Theory and Application, 2016, 9(5): 251 – 266.
[9] DENG Z, WANG Z, JIANG J. A new algorithm for fast mining frequent itemsets using N-lists [J]. SCIENCE CHINA Information Sciences, 2012, 55(9): 2008 – 2030.
[10] VO B, LE T, COENEN F, et al. Mining frequent itemsets using the N-list and subsume concepts [J]. International Journal of Machine Learning and Cybernetics, 2016, 7(2): 253 – 265.
[11] DAM T-L, LI K, FOURNIER-VIGER P, et al. An efficient algorithm for mining top-rank-*k* frequent patterns [J]. Applied Intelligence, 2016, 45(1): 96 – 111.
[12] MOHAMED M H, DARWIEESH M M. Efficient mining frequent itemsets algorithms [J]. International Journal of Machine Learning and Cybernetics, 2014, 5(6): 823 – 833.
[13] BURDICK D, CALIMLIM M, FLANNICK J, et al. MAFIA: a maximal frequent itemset algorithm [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1490 – 1504.

This work is partially supported by the Chongqing Postgraduate Research Innovation Fund (CYS15166).

LI Xiaolin, born in 1968, M. S., senior engineer. His research interests include mobile communication, big data.

DU Tuo, born in 1993, M. S. candidate. His research interests include big data, data mining.

LIU Biao, born in 1991, M. S. candidate. His research interests include distributed computing, data mining.