doi:10.11887/j.cn.202302006 http://journal.nudt.edu.cn

海量公交数据的人群画像算法*

张 锦[1,2],张建忠[1],汪 飞[3],郭 芊[1]

(1. 湖南师范大学 信息科学与工程学院,湖南 长沙 410006;2. 长沙理工大学 计算机与通信工程学院,湖南 长沙 410114;
3. 湖南师范大学 数学与统计学院,湖南 长沙 410006)

摘 要:面向海量公交数据的人群画像对分析城市群体出行特点、交通态势等极具价值,但对数据的处理存在耗时高、质量低、解释难等问题。提出一种海量公交数据人群画像的系统化解决策略,基于PageRank算法筛选出经过重要站点的人群轨迹,极大减少目标人群的轨迹数据;提出轨迹文本化分析方法来提高人群画像的可解释性;分析确定基于余弦距离的*K*-means算法作为人群画像分类的聚类算法。该算法在3 000万乘客公交出行数据上的实验表明:提出的解决策略能够较为系统性地解决海量公交数据的人群画像问题,同时基于余弦距离的*K*-means算法的聚类效果最好且准确率约达80%。将人群画像及其轨迹使用Flow Map进行可视化展示,结果符合真实世界的人群行为特征。

关键词:人群画像;PageRank算法;轨迹文本化;文本聚类

中图分类号:TP3-05 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:

文章编号:1001-2486(2023)02-055-10

Crowd profiling algorithm mass transit data

ZHANG Jin[1,2], *ZHANG Jianzhong*[1], *WANG Fei*[3], *GUO Qian*[1]

(1. College of Information Science and Engineering, Hunan Normal University, Changsha 410006, China;
2. School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China;
3. School of Mathematics and Statistics, Hunan Normal University, Changsha 410006, China)

Abstract: Crowd profiling of massive transit data is valuable for analyzing the travel characteristics and traffic trends of urban groups, but the processing of the data is time-consuming, low-quality and difficult to interpret. A systematic solution for crowd profiling of massive public transport data was proposed. Based on the PageRank algorithm, the trajectories of people passing through important stations were filtered out, which greatly reduced the trajectory data of the target population. A textual analysis method for trajectories was proposed to improve the interpretability of crowd profiling. And the *K*-means algorithm based on cosine distance as the clustering algorithm for crowd profiling was analysed and determined. The experiments on 30 million passengers' transit data show that the proposed algorithm can solve the problem of crowd profiling in massive transit data in a more systematic way, while the *K*-means algorithm based on cosine distance has the best clustering effect and the accuracy rate is about 80%. The crowd profiling and its trajectory were visually displayed by using Flow Map, and the results are consistent with real-world crowd behavioural characteristics.

Keywords: crowd portraits; PageRank algorithm; trajectory textualization; text clustering

现代公共交通技术利用先进的公交刷卡收费系统和公交刷卡信息数据库,记录着上百万条公交出行数据。研究发现,充分挖掘和利用公交乘客的刷卡数据,可以准确地分析出个人或群体的活动规律[1]。这些规律有助于公交路线规划和车辆调度,也可支持城市规划、社会行为分析等多个领域。

数据包括时空信息(上下车坐标、时间)和公交卡类型(成年卡、学生卡、老年卡)。存在三个问题:数据规模大,用聚类算法耗时长;公交轨迹点稀疏,聚类质量差;仅基于轨迹点聚类难以描述人群画像。

多数对公交数据的研究只关注识别或预测活动地点和出行方式,揭示人们一天中的出发和停留地点[2]。乘客出行特征涵盖时间、地点、目的地和时长等方面,但由于国内城市公交刷卡信息

* **收稿日期**:2021-02-26
基金项目:国家部委基金资助项目(31511010105);湖南省自然科学基金资助项目(2021JJ30456)
作者简介:张锦(1979—),男,河南信阳人,教授,博士,博士生导师,E-mail:jinzhang@hunnu.edu.cn;
汪飞(通信作者),男,安徽枞阳人,讲师,博士,E-mail:wangfei@hunnu.edu.cn

缺少持卡者类型描述,研究仅限于全客流出行模式特征的分析,无法描述不同年龄段的特征和挖掘不同人群的活跃模式[3]。文献[4]提出了由公交通勤出行模式的出行链提取的“四阶段法”。文献[5]结合(point of interest,POI)数据探讨了乘客出行功能区分布规律。文献[6]通过隐含狄利克雷分布(latent Dirichlet allocation, LDA)模型对每个热点区域不同时间段上下客流量进行分析,发掘乘客热点功能区域。

此外,轨迹文本化是将轨迹以文本形式呈现,方便了解每条轨迹的出行地区属性。文献[7]提出了一种新的把轨迹数据转化成文本的形式,用适当的特征去描述轨迹,十分依赖文本化时提取的特征。文献[8]利用 bigram 主题模型提取轨迹主题,并设计了多个链接视图的视觉分析系统。文献[9]提出基于用户和 POI 概况的旅游推荐系统下使用的推荐模型和算法。文献[10]利用 POI 等动态数据建立模型来描述空间使用率与车站区域的其他特征之间的关系。文献[11]提出基于重力的模型来估计中国上海市中心区域通勤模式。文献[12]以手机基站位置划分城市单元块,将 POI 与聚类结果的重叠率实现区域划分。文献[13]通过量化分析街区 POI 密度分值,进行武汉市核心城区功能分区。

另外,轨迹聚类是指将轨迹数据集划分成若干个子集的过程,每个子集为一个簇,使得每个簇内的轨迹彼此高度相似[14]。其目的就是挖掘轨迹大数据的移动模式,通过对聚类结果分析得到移动对象的出行规律。文献[15]提出了一种初始点优化与参数自适应的改进算法,优先对高密度簇进行聚类,即能对变化密度的数据集进行聚类。文献[16]提出了一种基于 MFTSM 的轨迹聚类算法,利用基于区域计算的位置距离来解决轨迹的连续性问题。文献[17]重新定义轨迹核心距离与轨迹可达距离,用邻接表代替空间索引来降低算法的复杂度。

针对以上问题,结合上述学者的研究,本文根据公交出行轨迹的相似性来分析人群轨迹特征,再使用自然语言描述人群轨迹的特征,从而可以更加清晰地了解人群出行规律,描绘出人群画像,同时也能进一步挖掘不同人群(不同年龄段的乘客)在城市各个区域的隐藏活跃模式。

此外,本文基于 PageRank 算法[18]提出重点地区人群筛选方法,提取出行次数多、去热点地区多的乘客轨迹数据,减少非重点地区的轨迹数据,从而减少数据量、提高处理效率。同时,按年龄段和工作日或休息日划分乘客轨迹数据,串联每位乘客的轨迹,提升数据质量和后续聚类算法的结果。将划分后的轨迹数据集与新加坡 POI[19]数据融合,文本化表现每位乘客的轨迹,并使用文本聚类算法对人群文本轨迹进行分类,以得到易于解释的轨迹类别特征,即人群画像。

1 算法构建

1.1 算法流程

本文的算法流程如图 1 所示。首先,将交通数据进行预处理,随后使用基于 PageRank 算法的人群筛选方法,从而减少数据量。再将筛选后的乘客刷卡数据串联起来,形成完整的轨迹数据。对于 POI 数据,首先将该数据预处理并将这些数据重新划分出 15 种功能性数据。随后将该数据与轨迹数据相结合,形成文本轨迹数据。最后通过使用聚类算法得到人群画像。

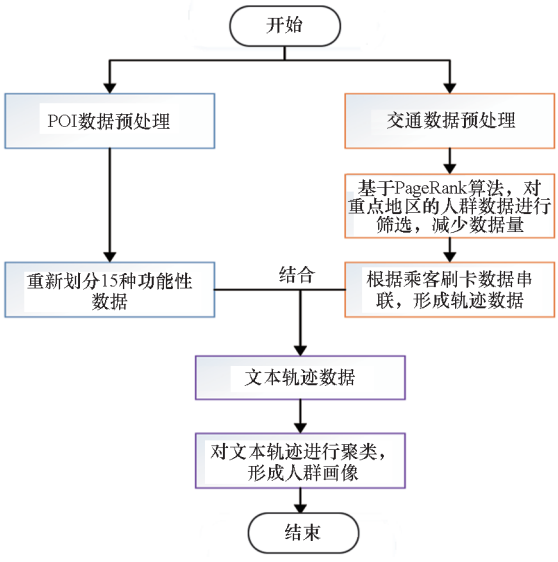


图 1 算法流程

Fig. 1 Algorithm flow chart

1.2 数据预处理

1.2.1 公交出行数据预处理

因为部分公交出行数据存在某些数据为空或数据字段偏移的情况,所以需要对数据进行整理。在对存在某些数据为空或数据字段偏移的数据进行数据规范化处理,并依据上下文补充数据和删除无法获取信息的数据后,得到了规范的公交出行数据。由于出行数据中原始 3 类人群(成年人、老年人和学生)出行方式差异较大,为便于更准确地描述人群画像,之后做了如下处理:

1)截取清洗后的新加坡公交车连续一周内所有的公交车刷卡数据,并且将这些数据按照工

作日和休息日进行划分。

2）将工作日和休息日的数据按照乘客年龄属性划分,得到的 6 组数据分别为:成年人工作日的刷卡数据、成年人休息日的刷卡数据、老年人工作日的刷卡数据、老年人休息日的刷卡数据、学生工作日的刷卡数据和学生休息日的刷卡数据。

3）将这 6 组数据中每日都有乘车记录且乘车次数至少为 2 的乘客数据筛选出来,最终得到了约 44 万名成年人乘客的刷卡数据、约 6 万名老年人乘客的刷卡数据和约 4 万名学生乘客的刷卡数据。

1.2.2　POI 数据预处理

因为 POI 数据为英文数据,这些数据中存在描述地点相同,但是字母大小写不一致的数据,所以首先将 POI 数据字体变为小写字母,然后删除数据中重复、指向不明和无效的数据。此外,由于 POI 数据的功能性指向过多,不利于后续工作的进行,将这些 POI 数据按功能性进行重新划分,把功能性相近的 POI 设置新的功能性,最后得到 15 种功能性类别。这 15 类功能性分别为餐饮、商业零售、服务行业、公共服务、休闲娱乐、居住、教育、宗教场所、医疗、景点、金融、政府机构、交通站点、体育健身、公司企业。

1.3　人群筛选

由于新加坡公交出行数据规模极大,此数据中每天有上百万条刷卡记录。如果直接将聚类算法使用在该数据中,会使聚类算法的时间消耗过长。为解决该问题,提出运用 PageRank 算法的重点地区人群筛选方法,通过该方法提取出行次数多且去热点地区次数多的乘客轨迹数据,极大地减少了非重点地区的人群轨迹数据,从而减少数据量和提高数据处理效率。

PageRank 算法,又称网页排名算法,是一种由搜索引擎根据网页之间的超链接计算的技术,用来体现网页的相关性和重要性。该算法的主要计算过程如图 2 所示。令 A 至 D 四个网页的初始重要性值为 1,再将每个网页的值除以其网页的出度 c,即得到每个网页之间链接的贡献度;最后对指向每个网页链接的贡献度求和,得到每个网页的重要性的值。

不同于传统 PageRank 的目的在于计算网页的重要性,本文方法的目的在于计算用户轨迹的重要性。具体包含两个主要过程:

1）根据各公交站点用户上下车频率,计算该站点的重要性数值;

2）根据用户经过的公交站点的重要性数值,计算用户轨迹的贡献率。

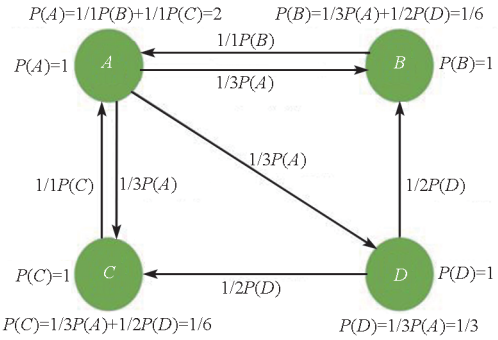


图 2　PageRank 算法的主要计算过程

Fig. 2　Main calculation process of the PageRank algorithm

站点重要性数值的计算公式为:

$$P(A)=(1-d)+d\left[\frac{P(T_1)}{C(T_1)}+\cdots+\frac{P(T_i)}{C(T_i)}+\cdots+\frac{P(T_n)}{C(T_n)}\right] \tag{1}$$

式中,$P(A)$是地点 A 的重要性数值;T_i 是所有指向地点 A 的路径的第 i 条路径;$P(T_i)$是地点 T_i 的重要性数值;$C(T_i)$是地点 T_i 的出度,也就是 T_i 指向其他地点的边的个数;d 为阻尼系数,表示在任意时刻人们到达某地区后继续出行的概率。在对所有的公交站点进行重要性数值计算后,每个公交站点都生成一个重要性数值,且用户上下车越多的站点,其重要性数值越大。

考虑到,如果用户经常去重要性数值较大的公交站点,那么与该用户轨迹相似的用户轨迹更多,从而可以认为该用户的轨迹对于计算用户画像的贡献度也越高。对每组数据设置一个阈值,用户的轨迹数据贡献率超过该阈值的,是有效的轨迹数据。用户轨迹贡献率的具体步骤如下:

1）将新加坡成年人工作日数据中的所有公交站点 N_{all}和出行的轨迹代入式(1),然后得到该组数据的所有出行公交站点对应的重要性数值 $P(N_{\text{all}})$。

2）根据该组人群中每名乘客 u 出行经过的公交站点 n 计算该名乘客轨迹的重要性数值之和,即 $S(u)=\sum_{i\in n}P(i)$。

3）对该组数据经过调试设置一个阈值 T,只有每名乘客轨迹的重要性数值之和大于该阈值,即 $S(u)>T$ 时,才将该乘客的轨迹数据保存下来。保存下来的轨迹数据即为本组数据中轨迹贡献率高的乘客数据。

4）将剩下 5 组数据(成年人休息日数据、老年人工作日数据、老年人休息日数据、学生工作日

数据、学生休息日数据）重复步骤 1 ~ 3，最后共得到 6 组轨迹贡献度高的乘客数据集。

1.4 轨迹文本化及聚类

文本分析具有良好的可解释性，而且存在有效的分析方法。在自然语言处理的领域，常用的文本处理方法通常是将文本进行分词与清洗，从而获取关键词语，再将文档嵌入词袋模型或者词向量模型，从而获得合适且表达能力强的特征。这些特征可以直接被机器学习模型或者深度学习模型使用以进行聚类或者分类等。聚类算法是一种无监督的机器学习方法，由于不需要预先对数据进行手工文档的标注，因此该方法具有较高的自动化处理能力。

经过人群筛选之后，将 6 组轨迹贡献度高的乘客轨迹数据集与 POI 数据相融合，得到文本化的轨迹数据。通过将每位乘客的文本轨迹数据导入词频 - 逆文本频率（term frequency-inverse document frequency，TF-IDF）算法中进行计算，以乘客轨迹作为文档，而 POI 作为关键词得到每位乘客的文本轨迹数据关键词的 TF-IDF 值。在此基础上，采用 *K*-means 算法进行聚类，并且比较了使用两种不同距离度量下的聚类结果。将两种聚类算法的结果使用 T 分布随机邻居嵌入（T-distributed stochastic neighbor embedding，T-SEN）算法[20]进行数据降维，利用散点图来展示两种算法的聚类效果，从而直观地比较这两种聚类算法的优劣性。

1.4.1 轨迹文本化

当确定一个乘客轨迹的坐标点时，以该坐标点为中心，计算该点周围 500 m 区域的经纬度，然后将所有的 POI 数据中属于该区域经纬度范围的 POI 属性数据提取出来并确定该坐标所属的功能性。在确定该坐标所属的功能性时，如果仅将该坐标点内数量最多的属性设置为该点的功能性，可能会导致结果存在较大的误差。因此使用文献[21]的方法对 POI 数据进行加权计算：

1）将这 15 类 POI 数据中每一类的数量 N_i 进行统计，在将所有的 POI 数量 N_{all} 除以 N_i，分别得到该类的权重 W_i，即：

$$W_i = \frac{N_{\text{all}}}{N_i} \tag{2}$$

2）对于坐标点范围内的所有 POI 数据，分别按照这 15 类的数量 n_i 进行统计。再将 n_i 乘以该类的权重 W_i 后除以该范围内所有的 POI 数量 n_{all}。最终得到该范围内的每类功能性的概率 P_i，即：

$$P_i = \frac{n_i}{n_{\text{all}}} \times W_i \tag{3}$$

3）因为某些地区存在很多不同功能性的 POI，随着时间的流逝该地区的主要功能性可能会发生变化，所以对这 15 种 POI 类型分别设置一个时间变化数 O_i，最后得到：

$$P_i = \frac{n_i}{n_{\text{all}}} \times W_i + O_i \tag{4}$$

4）将每个坐标点的上下车时间与上下车地点代入式（4）进行计算，从而分别得到该地区的各个功能性的概率值。随后选择该地概率值最高的两个功能性（不重复且 P_i 都大于 0）作为该时刻和该地点的功能性。最后，将每个乘客的所有轨迹点串联起来，得到每个乘客的文本轨迹。

1.4.2 文本聚类

TF-IDF[22]是一种用于信息检索与文本挖掘的常用加权技术和统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。在该方法中，TF 表示的是关键词在文本出现的频率，即：

$$T_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{5}$$

式中，$n_{i,j}$ 是词语 i 在文件 d 中出现的次数，$\sum_k n_{k,j}$ 表示文件 d 中所有词语出现的次数总和。

而 IDF 表示的是逆向文件频率，即：

$$I_i = \lg \frac{|D|}{1 + |\{j : t_i \in d_j\}|} \tag{6}$$

式中，$|D|$ 表示所有文本的数量，$|\{j : t_i \in d_j\}|$ 表示包含词语 t_i 的文件数目。TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类，即：

$$V_{\text{TF-IDF}} = T_{i,j} \cdot I_i \tag{7}$$

因此如果 TF-IDF 的值 $V_{\text{TF-IDF}}$ 越大，则表示该特征词对这个文本的重要性越大。

由于每个乘客的轨迹即是一个文本，而且所有文本由功能性关键词构成，因此无须划分关键词，只需要将每个功能性词语进行词频统计，再代入式（5）~（7）中，即可得到该乘客文本轨迹中各个功能性关键词的 TF-IDF 值并将其储存下来。将成年人工作日、成年人休息日、老年人工作日、老年人休息日、学生工作日、学生休息日这 6 组轨

迹贡献度高的乘客轨迹数据分别使用基于余弦距离的 *K*-means 聚类算法和基于欧氏距离的 *K*-means 聚类算法进行聚类运算,使用 T-SNE 算法对两种聚类算法的结果进行数据降维,使用散点图来展示两种算法的聚类效果。

2　实验及结果

2.1　数据描述

使用2012 年 12 月 5 日至 11 日连续 7 天非假期的新加坡公交出行数据进行研究,共3 000万条公交车和地铁的刷卡数据。因为乘客乘车时,上下车都需要刷一次 IC 卡来记录和支付,所以该数据完整地记录了每位乘客的 ID、上下车刷卡时间、上下车地点的经纬度。此外每位乘客的年龄属性(成年人、老年人和学生)也被记录在该数据中。根据这些数据,可以得到不同年龄段每名乘客的出行轨迹数据。在地理信息系统中,一个 POI 可以是一栋房子、一个商铺、一个公交站等,且一条完整 POI 数据必须包含该地点的名称、功能性、经纬度等数据[12]。通过调用 Google Map 的应用程序编程接口(application programming interface, API)获取新加坡的 POI 数据,最后共得到 4 万条数据。

2.2　贡献度高的乘客数据提取

经过轨迹文本化处理后,一共得到了约 44 万名成年人乘客的轨迹数据、约 5 万名老年人乘客的轨迹数据和约 3.1 万名学生乘客的轨迹数据。把这些数据使用 PageRank 算法后,得到了 4 529 个公交站点以及这些站点的重要性数值。随后使用 1.2 节的方法处理成年人工作日、成年人休息日、老年人工作日、老年人休息日、学生工作日、学生休息日这 6 组数据,并且得到了这 6 组数据中每位乘客的轨迹重要性数值之和。对阈值 T 进行尝试性设置,得到在不同的阈值 T 下每组数据的乘客数量以及经过的公交站点数量,如表 1 所示。设定的 T 必须满足以下两个条件:一是乘客数量尽可能少;二是公交站点数量尽可能多。只有满足以上两个条件的 T 所对应的数据,才能够保证在数据量变小的情况下,对后续结果的质量影响较小。因此这 6 组数据的阈值分别确定为 0.040(成年人工作日)、0.035(成年人休息日)、0.015(老年人工作日)、0.015(老年人休息日)、0.020(学生工作日)、0.020(学生休息日)。

表 1　在不同的阈值 T 下每组数据的乘客人数与公交站点数量

Tab. 1　Number of passengers and bus stops in each group of data under different T

数据类型		T							
		0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045
成年人工作日	乘客	297 311	212 022	143 128	91 455	56 798	33 919	19 402	10 750
	公交站点	4 326	4 231	4 071	3 907	3 828	3 607	3 572	2 802
成年人休息日	乘客	233 328	152 470	92 037	52 542	28 837	14 859	7 195	3 319
	公交站点	4 405	4 372	3 978	3 665	3 476	3 269	2 751	2 274
老年人休息日	乘客	15 696	6 818	2 698	970	338	108	30	5
	公交站点	4 103	3 963	1 701	1 206	402	98	70	11
老年人休息日	乘客	13 547	3582	712	102	14	1	0	0
	公交站点	3 570	2 878	206	97	5	4	0	0
学生工作日	乘客	18 108	11 712	6 763	3 458	1 558	624	211	72
	公交站点	4 328	4 328	4 307	2 741	1 504	402	107	32
学生工作日	乘客	22 698	14 646	8 608	4 458	2 183	985	408	142
	公交站点	4 032	3 875	3 667	2 560	1 795	601	432	90

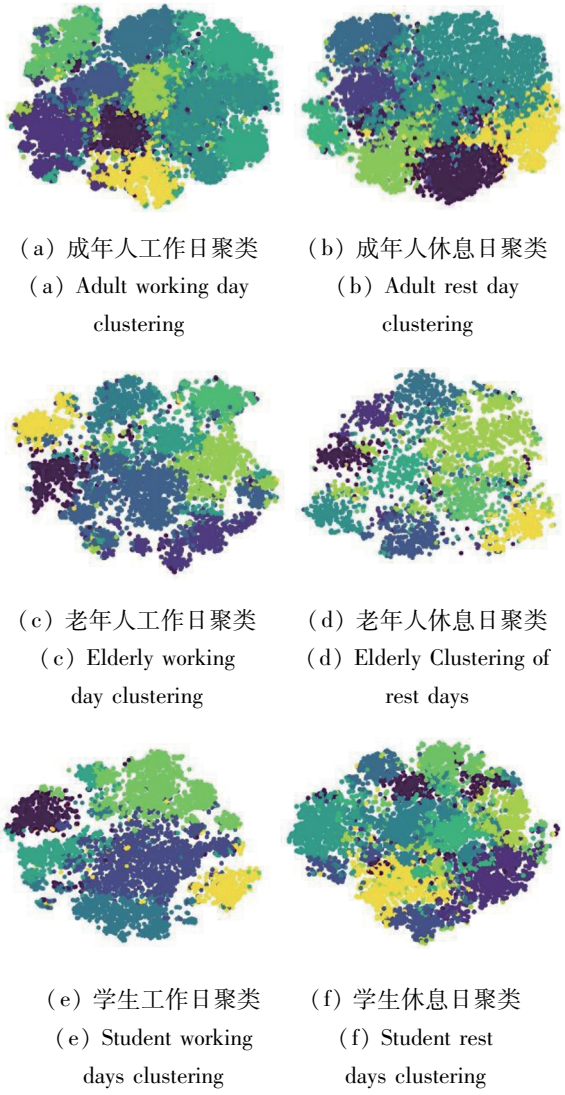
2.3　文本轨迹聚类

根据 1.4 节所述,先将这 6 组数据的乘客轨迹文本化,再将处理后的数据分别使用基于余弦距离的 *K*-means 聚类算法和基于欧氏距离的 *K*-means 聚类算法进行计算,接着将结果使用 T-SNE算法分别进行数据降维并使用散点图展示其聚类效果。然后从每组轨迹贡献度高的乘客数据中选取 1 000 条轨迹数据,将这些数据经

过 TF-IDF 处理后进行标记。最后标记的数据分别与基于余弦距离的聚类结果和基于欧氏距离的聚类结果进行比对检验,并对准确性进行计算。

2.3.1 聚类结果对比

将处理后的数据使用基于余弦距离的 *K*-means 聚类,再将结果使用 T-SNE 算法进行数据降维并展示其聚类效果,结果如图 3 所示。使用基于欧氏距离的聚类结果如图 4 所示。图中颜色分布越集中,说明聚类效果越好。



(a) 成年人工作日聚类 (a) Adult working day clustering

(b) 成年人休息日聚类 (b) Adult rest day clustering

(c) 老年人工作日聚类 (c) Elderly working day clustering

(d) 老年人休息日聚类 (d) Elderly Clustering of rest days

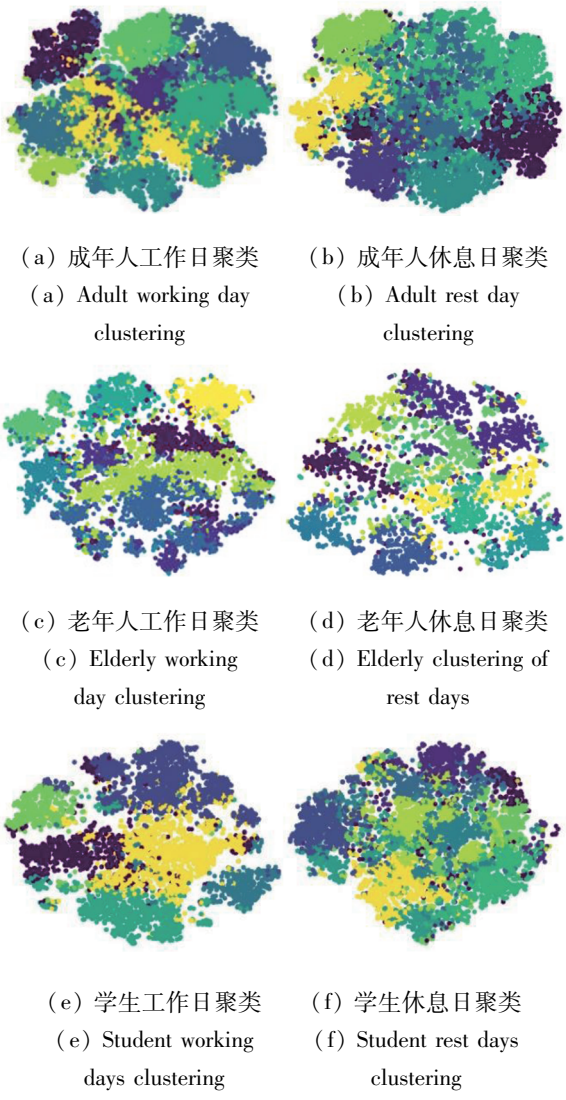
(e) 学生工作日聚类 (e) Student working days clustering

(f) 学生休息日聚类 (f) Student rest days clustering

图 3 基于余弦距离的 *K*-means 聚类算法的结果

Fig. 3 Results of *K*-means clustering algorithm based on cosine distance

通过图 3 与图 4 的对比可以明显看出,基于余弦距离的 *K*-means 聚类算法得到的簇的分布要优于基于欧氏距离的 *K*-means 聚类算法得到的簇的分布。对基于余弦距离的 *K*-means 聚类算法的结果进行归整,将同一类型乘客的轨迹数据提取出来,再使用一次 TF-IDF 算法,得到排名前 4 的关键词,即该类人群常去的功能区域。



(a) 成年人工作日聚类 (a) Adult working day clustering

(b) 成年人休息日聚类 (b) Adult rest day clustering

(c) 老年人工作日聚类 (c) Elderly working day clustering

(d) 老年人休息日聚类 (d) Elderly clustering of rest days

(e) 学生工作日聚类 (e) Student working days clustering

(f) 学生休息日聚类 (f) Student rest days clustering

图 4 基于欧氏距离的 *K*-means 聚类算法的结果

Fig. 4 Results of *K*-means clustering algorithm based on Euclidean distance

1) 对成年人工作日轨迹贡献度高的乘客数据,聚类生成了 9 类乘客轨迹。每类乘客出行常去的功能区域如表 2 所示。

2) 对于成年人休息日轨迹贡献度高的乘客数据,聚类生成了 7 类人群轨迹,每类乘客出行常去的功能区域如表 3 所示。

3) 对于老年人工作日轨迹贡献度高的乘客数据,聚类生成了 8 类人群轨迹,每类乘客出行常去的功能区域如表 4 所示。

4) 对于老年人休息日轨迹贡献度高的乘客数据,聚类生成了 9 类人群轨迹,每类乘客出行常去的功能区域如表 5 所示。

5) 对于学生工作日轨迹贡献度高的乘客数据,聚类生成了 6 类人群轨迹,每类乘客出行常去的功能区域如表 6 所示。

表 2　成年人工作日轨迹贡献度高的乘客聚类结果

Tab. 2　Passenger clustering results with high contribution of adult trajectory on workday

常去的功能区域	数量
居住,餐饮,公司企业,政府机构	1 167
居住,餐饮,商业零售,教育	2 648
居住,餐饮,公司企业,宗教场所	1 360
居住,餐饮,公司企业,医疗	1 979
居住,餐饮,公司企业,商业零售	4 183
居住,餐饮,公司企业,金融	3 542
居住,餐饮,公司企业,交通站点	1 566
居住,餐饮,公司企业,服务行业	1 487
居住,餐饮,公司企业,体育健身	1 470

表 3　成年人休息日轨迹贡献度高的乘客聚类结果

Tab. 3　Passenger clustering results with high contribution of adult trajectory on weekend

常去的功能区域	数量
居住,餐饮,商业零售,医疗	2 118
居住,餐饮,商业零售,交通站点	1 263
居住,餐饮,商业零售,教育	1 374
居住,餐饮,商业零售,休闲娱乐	6 103
居住,餐饮,商业零售,宗教场所	751
居住,餐饮,商业零售,体育健身	1 638
居住,餐饮,商业零售,公司企业	1 612

表 4　老年人工作日轨迹贡献度高的乘客聚类结果

Tab. 4　Passenger clustering results with high contribution of senior trajectory on workday

常去的功能区域	数量
居住,餐饮,商业零售,体育健身	593
居住,餐饮,休闲娱乐,公司企业	781
居住,餐饮,商业零售,休闲娱乐	2 099
居住,餐饮,商业零售,交通站点	676
居住,餐饮,商业零售,服务行业	660
居住,餐饮,商业零售,教育	520
居住,餐饮,商业零售,医疗	995
居住,餐饮,商业零售,宗教场所	494

6) 对于学生休息日轨迹贡献度高的乘客数据,聚类生成了 10 类人群轨迹,每类乘客出行常去的功能区域如表 7 所示。

表 5　老年人休息日轨迹贡献度高的乘客聚类结果

Tab. 5　Passenger clustering results with high contribution of senior trajectory on weekend

常去的功能区域	数量
居住,餐饮,公司企业,政府机构	248
居住,餐饮,商业零售,金融	249
居住,餐饮,商业零售,交通站点	383
居住,餐饮,商业零售,医疗	459
居住,餐饮,商业零售,教育	246
居住,餐饮,商业零售,服务行业	365
居住,餐饮,商业零售,体育健身	343
居住,餐饮,商业零售,休闲娱乐	1 027
居住,餐饮,商业零售,宗教场所	262

表 6　学生工作日轨迹贡献度高的乘客聚类结果

Tab. 6　Passenger clustering results with high contribution of student trajectory on workday

常去的功能区域	数量
教育,居住,餐饮,宗教场所	588
教育,居住,餐饮,服务行业	2 065
教育,居住,餐饮,体育健身	1 135
教育,居住,餐饮,医疗	793
教育,居住,餐饮,交通站点	1 645
教育,居住,餐饮,公司企业	537

表 7　学生休息日轨迹贡献度高的乘客聚类结果

Tab. 7　Passenger clustering results with high contribution of student trajectory on weekend

常去的功能区域	数量
居住,餐饮,商业零售,政府机构	604
居住,餐饮,商业零售,医疗	1 037
居住,餐饮,商业零售,金融	638
居住,餐饮,商业零售,宗教场所	415
居住,餐饮,商业零售,服务行业	1 431
居住,餐饮,商业零售,教育	1 015
居住,餐饮,商业零售,休闲娱乐	908
居住,餐饮,商业零售,体育健身	768
居住,餐饮,商业零售,公司企业	905
居住,餐饮,服务行业,交通站点	887

2.3.2　准确率对比

因为聚类算法是属于无监督的机器学习算法,所以该算法需要对原始数据进行标注,再与聚

类计算后的结果进行检验。因此做了以下步骤来验证准确率：

1）从上述6组轨迹贡献度高的乘客数据中，每组随机抽取1 000名乘客的轨迹数据，再使用TF-IDF算法进行计算，对每条轨迹取排名前4的关键词作为该名乘客的标签。

2）将每位乘客的标签与该乘客所对应类的常去功能区域进行对比。由于每组数据聚类结果的常去功能区域的前三个关键词是基本相同的，每一类的区别在于第4个关键词。因此，在对每位乘客的标签与该乘客对应的功能区域进行检验时，在第4个关键词必须存在的情况下，剩下3个关键词至少存在2个，则确定该乘客分类正确。对于第3个关键词与其他类的第3个关键词也不相同的情况，只有在第3和第4个关键词都存在的情况下，剩下2个关键词至少存在1个，才能确定该乘客分类正确。

3）将划为分类正确的乘客数量统计出来，再除以该组总人数，得到该组乘客的准确率。将所有分类正确的乘客数量统计出来，除以抽取的所有乘客轨迹数据，得到所有数据的准确率。

综上，本文将基于余弦距离的*K*-means聚类算法和基于欧氏距离的*K*-means聚类算法的结果准确率进行计算，如表8所示。

表8 基于余弦距离与欧氏距离的*K*-means聚类算法结果准确率

Tab. 8 Accuracy of *K*-means clustering algorithm based on cosine distance and Euclidean distance

%

数据类型	余弦距离	欧式距离
成年人工作日	75.50	58.20
成年人休息日	70.60	77.10
老年人工作日	81.90	52.00
老年人休息日	80.80	53.40
学生工作日	82.60	68.50
学生休息日	80.20	45.90
平均准确率	78.60	59.18

2.4 案例分析

为了更加直观地展示每类人群轨迹的区别，展现出人群画像的区别，以成年人工作日的数据进行案例分析，并分别将乘客的轨迹数据经处理导入Flow Map中展示。对于Flow Map生成的轨迹图，人群在两地流动越频繁，两地间的线段越粗；人群在某点聚集得越多，该点越大。

对于成年人工作日的数据，经过上述处理一共分为9类人群画像。根据表2可知，大部分成年人在工作日是在居住地、餐饮、公司企业和“第4个地方”活动。而这“第4个地方”正是区分每类乘客的关键，这与文献[3]所述相似。将这9类乘客的轨迹数据导入Flow Map中，可以得到成年人工作日出行的主要流动情况，如图5所示。可以清楚地了解不同轨迹出行的乘客常去的地点以及出行的频率。从图中可知，因为新加坡的南部是政治、经济和文化活动的中心地区且基础设施完善，所以新加坡南部的居住地、餐饮业、政府机构、宗教场所、商业零售、金融机构、服务行业、体育健身场所和公司企业在这里聚集较多；教育场所（小学、中学）主要分布新加坡北部和南部地区，东西地区分布较少，因此许多成年人会经常去往新加坡北部地区和南部地区；而对于医疗场所来说，由于新加坡大型的公立医院分布在新加坡的东部地区和西部地区，因此当人们需要去医疗场所时会常前往这两个地区。

（a）居住、餐饮、公司企业、政府机构

（a）Residential, catering, companies, government institutions

（b）居住、餐饮、商业零售、教育

（b）Residential, catering, retail, education

（c）居住、餐饮、公司企业、宗教场所

（c）Residential, catering, companies, religious places

(d) 居住、餐饮、公司企业、医疗
(d) Residential, catering, companies, medical

(e) 居住、餐饮、公司企业、商业零售
(e) Residential, catering, companies, commercial retail

(f) 居住、餐饮、公司企业、金融
(f) Residential, catering, companies, financial

(g) 居住、餐饮、公司企业、交通站点
(g) Residential, catering, companies, transportation stations

(h) 居住、餐饮、公司企业、服务行业
(h) Residential, catering, companies, service

(i) 居住、餐饮、公司企业、体育健身
(i) Residential, catering, companies, sports and fitness

图 5　成年人工作日贡献度高的乘客聚类特性及轨迹图
Fig. 5　Clustering characteristics and trajectory of adult passengers with high contribution from working days

3　结论

1)对新加坡乘客的出行轨迹使用基于 PageRank 算法的重点地区人群筛选方法,提取出行次数多且去热点地区次数多的乘客轨迹数据,极大地减少了非热点地区的乘公交频率较少的人群轨迹数据,同时提高数据处理效率。

2)将筛选后的乘客轨迹数据按照年龄段与一周内连续的工作日与休息日进行划分,并将每位乘客的轨迹数据串联起来,形成完整的轨迹数据集,从而提升数据质量,为提升后续聚类算法的结果质量提供基础。

3)将划分后的轨迹数据集与新加坡 POI 数据相融合得到每位乘客的文本化轨迹,然后使用 TF-IDF 算法对文本轨迹的关键词进行提取。

4)分别使用基于余弦距离的 *K*-means 算法与基于欧氏距离的 *K*-means 算法对上述关键词进行聚类,并对产生的结果进行对比。结果表示基于余弦距离的 *K*-means 算法对乘客轨迹的聚类效果更好,该算法的准确率接近 80% 且更稳定。

5)将分类结果使用 Flow Map 进行可视化展示,并对每类人群的画像进行简单的分析。通过上述工作,可为城市规划、社会行为分析等多个应用领域提供数据支撑,方便城市资源的合理调度与建设,对城市建设和发展做出最优决策。

参考文献(References)

[1] 赵明星. 基于乘客出行特征的公交线网优化方法研究[D]. 石家庄: 河北科技大学, 2022.
ZHAO M X. Optimization of bus route network based on passenger travel characteristics method research [D]. Shijiazhuang: Hebei University of Science and Technology, 2022. (in Chinese)

[2] LEGARA E F T, MONTEROLA C P. Inferring passenger type from commuter eigentravel matrices[J]. Transportmetrica B: Transport Dynamics, 2018, 6(3): 230 – 250.

[3] 王长硕，蒲英霞. 基于 Labeled-LDA 模型的居民群体分类与出行特征分析[J]. 计算机应用与软件，2022，39(11)：17－24.
WANG C S，PU Y X. Analysis of classification and activity characteristics of urban residents based on Labeled-LDA model[J]. Computer Applications and Software，2022，39(11)：17－24.（in Chinese）

[4] 王月玥. 基于多源数据的公共交通通勤出行特征提取方法研究[D]. 北京：北京工业大学，2014.
WANG Y Y. Research on methods of extracting commuting trip characteristic based on public transportation multi-source data[D]. Beijing：Beijing University of Technology，2014.（in Chinese）

[5] 程静，刘家骏，高勇. 基于时间序列聚类方法分析北京出租车出行量的时空特征[J]. 地球信息科学学报，2016，18(9)：1227－1239.
CHENG J，LIU J J，GAO Y. Analyzing the spatio-temporal characteristics of Beijing′s OD trip volume based on time series clustering method[J]. Journal of Geo-Information Science，2016，18(9)：1227－1239.（in Chinese）

[6] 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.
SUN G D，ZHANG B，LIU Y Q，et al. Taxi hot area function discovery based on passenger data[J]. Computer Engineering，2017，43(5)：16－22.（in Chinese）

[7] AL-DOHUKI S，WU Y Y，KAMW F，et al. SemanticTraj：a new approach to interacting with massive taxi trajectories[J]. IEEE Transactions on Visualization and Computer Graphics，2017，23(1)：11－20.

[8] LIU H，JIN S C，YAN Y Y，et al. Visual analytics of taxi trajectory data via topical sub-trajectories[J]. Visual Informatics，2019，3(3)：140－149.

[9] SANTOS F，ALMEIDA A，MARTINS C，et al. Using POI functionality and accessibility levels for delivering personalized tourism recommendations[J]. Computers，Environment and Urban Systems，2019，77：101173.

[10] YE Z N，CHEN Y H，ZHANG L. The analysis of space use around Shanghai metro stations using dynamic data from mobile applications[J]. Transportation Research Procedia，2017，25：3147－3160.

[11] LI M Y，KWAN M P，WANG F H，et al. Using points-of-interest data to estimate commuting patterns in central Shanghai，China[J]. Journal of Transport Geography，2018，72：201－210.

[12] 蒋云良，董墨萱，范婧，等. 基于 POI 数据的城市功能区识别方法研究[J]. 浙江师范大学学报(自然科学版)，2017，40(4)：398－405.
JIANG Y L，DONG M X，FAN J，et al. Research on identifying urban regions of different functions based on POI data[J]. Journal of Zhejiang Normal University（Natural Sciences），2017，40(4)：398－405.（in Chinese）

[13] 康雨豪，王玥瑶，夏竹君，等. 利用 POI 数据的武汉城市功能区划分与识别[J]. 测绘地理信息，2018，43(1)：81－85.
KANG Y H，WANG Y Y，XIA Z J，et al. Identification and classification of Wuhan urban districts based on POI[J]. Journal of Geomatics，2018，43(1)：81－85.（in Chinese）

[14] YUAN G，SUN P H，ZHAO J，et al. A review of moving object trajectory clustering algorithms[J]. Artificial Intelligence Review，2017，47(1)：123－144.

[15] DAI Y Y，LI C F，XU H. Density clustering algorithm with initial point optimization and parameter self-adaption[J]. Computer Engineering，2016，42(1)：203－209.

[16] YU Q Y，LUO Y L，CHEN C M，et al. Trajectory similarity clustering based on multi-feature distance measurement[J]. Applied Intelligence，2019，49(6)：2315－2338.

[17] 杨树亮，毕硕本，NKUNZIMANA A，等. 一种出租车载客轨迹空间聚类方法[J]. 计算机工程与应用，2018，54(14)：249－255.
YANG S L，BI S B，NKUNZIMANA A，et al. Spatial clustering method for taxi passenger trajectory[J]. Computer Engineering and Applications，2018，54(14)：249－255.（in Chinese）

[18] TORTOSA L，VICENT J F，YEGHIKYAN G. An algorithm for ranking the nodes of multiplex networks with data based on the PageRank concept[J]. Applied Mathematics and Computation，2021，392：125676.

[19] 张景奇，史文宝，修春亮. POI 数据在中国城市研究中的应用[J]. 地理科学，2021，41(1)：140－148.
ZHANG J Q，SHI W B，XIU C L. Urban research using points of interest data in China[J]. Scientia Geographica Sinica，2021，41(1)：140－148.（in Chinese）

[20] LAURENS V D M，HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research，2008，9(2605)：2579－2605.

[21] ZHANG J Z，WANG F，GUO Q，et al. Multisource data-driven visual analysis of urban crowd travel[J]. Journal of Physics：Conference Series，2021，1757(1)：012108.

[22] 罗燕，赵书良，李晓超，等. 基于词频统计的文本关键词提取方法[J]. 计算机应用，2016，36(3)：718－725.
LUO Y，ZHAO S L，LI X C，et al. Text keyword extraction method based on word frequency statistics[J]. Journal of Computer Applications，2016，36(3)：718－725.（in Chinese）