

引用格式: 初晨, 张恒才, 陆锋. 大型商场顾客消费行为轨迹推断[J]. 地球信息科学学报, 2022, 24(6): 1034-1046. [Chu C, Zhang H C, Lu F. Inferring consumption behavior of customers in shopping malls from indoor trajectories[J]. Journal of Geo-information Science, 2022, 24(6): 1034-1046.] DOI:10.12082/dqxxkx.2022.210690

大型商场顾客消费行为轨迹推断

初晨^{1,2}, 张恒才^{1,2*}, 陆锋^{1,2}

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101; 2. 中国科学院大学资源与环境学院, 北京 100049

Inferring Consumption Behavior of Customers in Shopping Malls from Indoor Trajectories

CHU Chen^{1,2}, ZHANG Hengcai^{1,2*}, LU Feng^{1,2}

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China, 2. College of Resource and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: How to obtain the consumption behavior of massive customers in large indoor shopping malls has always been a difficult problem in behavioral geography. However, with the explosive growth of indoor trajectory data in recent years, there's a great opportunity to solve this problem. Meanwhile, the lack of semantic information and poor data quality of indoor trajectory still pose challenges to the inference of consumer behavior. This study proposes a framework for customers' consumption behavior inference in shopping malls without collecting private personal consumption records. This framework integrates the Web text information of stores with movement features extracted from personal and historical customer trajectories. The semantic attributes of indoor stores are enhanced by introducing the crawled network text data of indoor stores, so as to realize the transformation from customer geometric trajectory to semantic trajectory. Specifically, the framework offers a method to model the customers' consumption feature from three aspects, including the raw trajectory's movement feature, semantic feature, and movement embedding feature. By employing the representation learning algorithm in extraction of customers' movement embedding feature, the framework can learn the movement pattern from the historical crowd trajectories and use the movement embedding feature to model movements of a single customer in a complex indoor environment automatically. Finally, the research realizes residents' consuming behavior inference by clustering the concatenated multi-sources consuming features and analyzing the clusters with statistic values and visualization. Through the experimental analysis of a real-world indoor trajectory dataset generated from a large shopping mall with 7045 customers, the inference result proves that the framework can effectively extract the spatial-temporal movement and consumption pattern of residents. Comparing with the classic feature extraction methods and typical clustering methods, the framework we propose achieves an improvement for up to 69.8% in the Silhouette Coefficient. This improvement illustrates that the customers' consumption behavior inferring framework we propose can identify the customers with

收稿日期: 2021-10-30; 修回日期: 2021-12-16.

基金项目: 国家重点研发计划项目(2021YFB3900803)。[**Foundation item:** National Key Research and Development Program of China, No.2021YFB3900803.]

作者简介: 初晨(1998—), 男, 山东青岛人, 硕士, 主要从事时空数据挖掘研究。E-mail: chuchen0411@igsrr.ac.cn

*通讯作者: 张恒才(1985—), 男, 山东济南人, 博士, 副研究员, 主要从事立体时空计算研究。E-mail: zhanghc@lreis.ac.cn

different consuming behaviors more effectively and cluster customers' feature with high dimension more precisely. Through the analysis of indoor customer clusters' movement pattern, the research finds out that the moving behavior of all shopping mall customers are affected directly and prominently by the design of indoor environment e.g., the distribution of functional zones, location of escalators, etc. Besides, the research also finds out that customers have strong preference to consume in the identical floor. The framework we proposed can identify customer groups with different consumption levels and movement patterns and discover consuming patterns from massive shopping mall customers without knowing their personal information. The application of the framework in inferring customer behavior patterns could provide a support for relative researches in behavioral geography.

Key words: indoor trajectory; trajectory data mining; indoor consumption behavior; trajectory clustering; crowd spatial-temporal movement; movement feature embedding; movement pattern mining; indoor space structure; human dynamic observation

*Corresponding author: ZHANG Hengcai, E-mail: zhanghc@lreis.ac.cn

摘要:如何获取大型商场内海量顾客消费行为一直是行为地理学面临的难点问题,而近年来爆发式增长的室内轨迹数据为这一问题解决提供了机遇,但室内轨迹的语义信息缺失、数据质量差等问题给推断顾客消费行为造成了挑战。本研究提出了一种顾及文本-轨迹的商场顾客消费行为轨迹推断框架,无需隐私敏感的顾客消费记录数据,可以获取大量顾客消费行为,该方法通过爬取室内店铺的网络文本,增强室内店铺语义属性,进而实现顾客几何轨迹到语义轨迹的转化提升,并引入了轨迹嵌入特征表示学习方法,捕捉群体轨迹之间的移动特征,综合轨迹移动特征、轨迹语义特征及顾客嵌入特征,通过高维聚类实现了大型商场顾客消费模式的推断。通过某大型商场7045位顾客的真实轨迹进行实验分析,实验结果表明,本文提出的方法与传统特征提取方法相比,聚类结果在轮廓系数上提升最高达69.8%,顾客消费行为提取准确率更高。研究发现,室内顾客移动具有一定楼层倾向性,并且室内空间结构如店铺位置、扶梯位置、功能区划分等,会影响顾客消费模式。本文提出的方法可以有效识别不同消费水平、移动特征的顾客群体,实现顾客消费行为的轨迹推断。

关键词:室内轨迹;轨迹挖掘;室内消费行为;轨迹聚类;人群时空行为;移动特征嵌入;移动模式挖掘;室内空间结构;人群动态观测

1 引言

消费行为是指在特定环境、社会、政治文化背景之下,作为空间决策主体消费者对商品、服务的购买,使用及消费等一系列时空行为^[1-3]。大型商场顾客购物选择与消费行为是商业需求的基本驱动力,与城市商业空间结构密切相关,能够为未来城市商业空间规划提供决策依据,消费时空行为研究是行为地理学热点研究问题^[4-6]。

目前已有消费行为研究集中在宏观尺度与中观尺度,如线上购物行为重塑顾客购物消费模式^[7],消费区位偏好提取^[8],消费行为与商业形态空间关系研究^[9],消费时空行为影响因素研究^[2]等。受限于目前主流的调查问卷获取数据方式存在的诸多弊端^[10],微观尺度商场顾客消费行为研究相对较少,一方面调查问卷耗时耗力且获取样本有限,另一方面调查问卷数据时效性不高,数据准确性、可用性较低。此外载有大量消费者消费行为的数据,如店铺会员、POS刷卡、网络购物、大规模电商数据等^[11-14],

由于隐私保护等原因无法大批量获取,造成目前消费时空行为研究难以在微观尺度开展。

近年来,室内定位技术蓬勃发展,如WIFI、iBeacon, UWB、伪卫星、音频等^[15-18]定位技术被逐步广泛应用,室内位置服务应用不断增多,爆发式增长的室内轨迹数据为大型商场顾客消费行为研究提供了机遇^[19]。目前已有大型商场顾客轨迹数据相关研究侧重于几何轨迹分析,如室内导航网络生成^[20]、室内导航路径设计算法的优化^[21]、室内聚集区域提取^[22]、室内频繁模式挖掘^[23]、室内位置预测^[24]等。在行为模式方面,已有研究集中在浅层移动行为,如通过语义轨迹的修改距离衡量轨迹相似度,进行移动行为的聚类研究^[25],通过分析特定室内场景下室内环境与人群移动行为间的联系,进行室内人群移动流挖掘^[26-27]等。但在深层语义挖掘,如消费行为、移动规律等方面研究仍然较少,Liu等^[28]尝试通过带有用户属性信息的行人轨迹,通过建立用户到访店铺与属性信息间的随机森林模型,推测顾客性别、年龄特性,描绘顾客画像,Yaeli等^[29]通过建

立室内商场的可视化数据分析平台进行商场顾客行为分析^[29]。室内轨迹数据的优势是数据量大,时效性高、连续性好,但利用商场顾客轨迹数据进行消费行为研究依然存在语义信息缺失,轨迹数据质量差,跳楼层严重的问题,而且由于室内定位轨迹往往都进行了脱密处理,不包含任何语义属性信息,难以推断顾客消费时空行为。

为解决上述问题,本文提出了一种顾及文本-轨迹的大型商场顾客消费行为推断学习框架,通过爬取室内环境店铺的网络文本数据,为顾客轨迹赋予语义属性,丰富顾客轨迹的消费语义信息;并引入表示学习方法,由群体轨迹数据提取个体顾客的移动特征,解决复杂地理环境下移动对象活动规律特征提取问题;通过对商场顾客进行基于消费特征的高维聚类分析,实现大型商场顾客不同类别消费行为模式提取与消费规律推断。

2 研究方法

2.1 研究思路

考虑到室内轨迹仅包含经纬度、时间等极少几何位置信息,但室内店铺的丰富语义属性信息大量存在于大众点评、百度地图等网络文本中,例如NIKE店铺为男装类店铺,平均消费价格为400元左右,本文通过爬取室内店铺的网络文本,提取顾客移动行为深层语义信息,基本思路如图1所示,通过将轨迹数据映射到店铺空间位置,利用网络文本提取顾客经过店铺序列的属性信息,获取属性信息丰富的顾客语义轨迹,开展商场顾客消费行为挖掘研究。

考虑到商场顾客在室内消费时移动行为的复杂性,本方法从轨迹移动特征、轨迹语义特征、顾客嵌入特征3个角度进行特征刻画。轨迹移动特征为原始轨迹中包含的速度、距离等移动信息,轨迹语义特征为通过网络文本丰富店铺信息之后,利用顾客经过的语义轨迹序列提取出的顾客消费特征,如平均消费水平等。轨迹移动特征和轨迹语义特征都是从单个顾客视角去构建特征,而室内由于顾客在店铺之间移动,店铺之间会形成相对固定的访问规律关系,本文引入表示学习方法,学习生成顾客轨迹的嵌入特征,作为顾客移动规律特征的体现。

图2为本文方法的基本流程图,首先通过对轨迹的分析提取了一系列顾客原始轨迹特征;而后通过Indoor-STDBSCAN算法^[24]对顾客的原始轨迹的轨迹点进行聚类,提取停留区域,进而生成代表顾客到访过店铺的语义轨迹,利用语义轨迹中的语义点包含的店铺文本信息,对访问该语义点的轨迹进行信息丰富,提取轨迹语义特征;而后通过表示学习方法,对各店铺在大量商场顾客历史轨迹中的店铺共现规律进行学习,获取商场内各店铺的特征向量,进而生成顾客移动嵌入特征。最后利用多元特征集成的消费行为特征,通过对顾客进行聚类分析,推断商场顾客室内消费行为。

2.2 室内顾客轨迹定义

定义1 轨迹点(Location Point, LP):顾客在室内活动时的定位数据点,如式(1)所示。

$$LP=(mac, time, x, y, floor) \quad (1)$$

式中: mac 为顾客接入互联网的物理地址,作为顾客的识别标识; $time$ 为获取到当前定位点的时间,

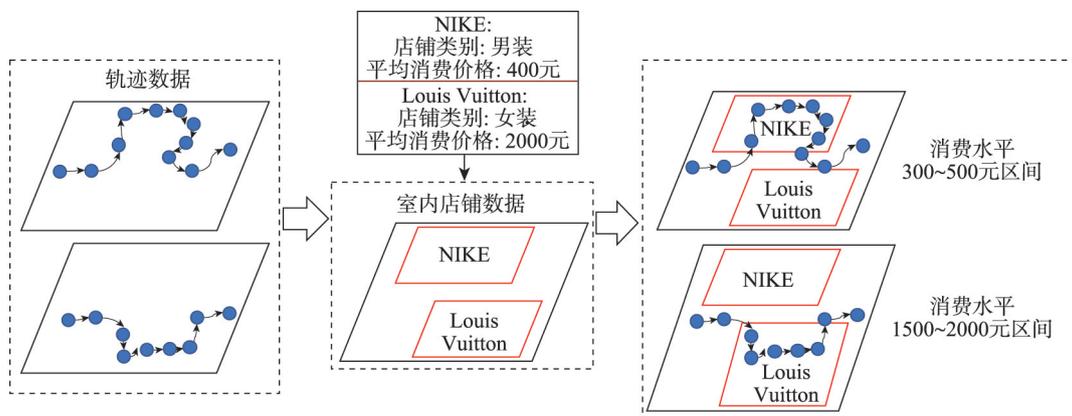


图1 基本思路示意图

Fig. 1 Schematic diagram of research's basic idea

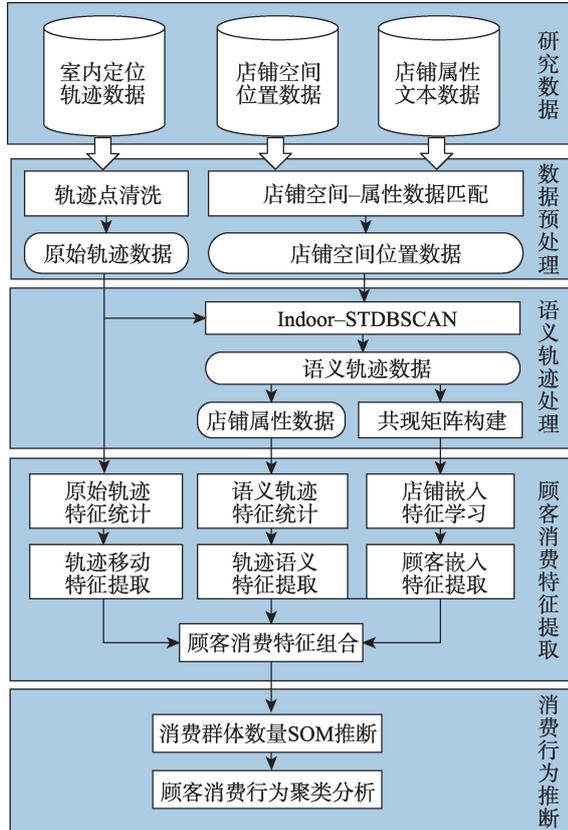


图2 研究技术路线示意

Fig. 2 Overview of research framework

x 、 y 表示顾客当前位置的地理坐标; $floor$ 为获取该定位点时顾客所在楼层。

定义2 原始轨迹(Raw Trajectory, RawTraj):由一系列定位点的集合组成的移动对象的运动状态表示。

$$RawTraj_{mac} = \{LP_i\}_{i=0}^n \quad (2)$$

式中: $RawTraj_{mac}$ 代表识别标识为 mac 的原始轨迹数据集; LP_i 为构成顾客移动轨迹的若干个定位地点,其中要求 $LP_{i+1}.time > LP_i.time$, 即轨迹中定位点按时间顺序排列。

定义3 停留点(Stay Point, SP):利用原始轨迹通过时空聚类算法提取的顾客在室内活动时停留的店铺。

$$SP = (POI, ArrT, LevT, Price) \quad (3)$$

式中: POI 为顾客经过的室内店铺; $ArrT$ 为顾客到达该店铺的时间; $LevT$ 为顾客离开该店铺的时间; $Price$ 为当前店铺的消费水平表示。

定义4 语义轨迹(Semantic Trajectory, SemanTraj):由顾客经过的一系列具体店铺构成。

$$SemanTraj_{mac} = \{SP_i\}_{i=0}^n \quad (4)$$

式中: SP_i 为顾客经过的第 i 个语义点,与原始轨迹相同,同样要求 $SP_{i+1}.ArrT > SP_i.LevT$, 即语义轨迹为按照顾客停留点生成的时间顺序排列成的序列。

2.3 顾客消费行为轨迹推断方法

2.3.1 轨迹移动特征提取

顾客的原始轨迹中包含最直接、最显著的顾客移动特征,这些特征在表示顾客移动模式时变量解释性好。为提取顾客在室内的移动行为,本文选取了表1所示的轨迹统计量作为顾客原始轨迹的特征表示,分别从时间、水平空间及室内环境中分异最明显的垂直空间上对商场顾客移动特征进行了表达。

表1 轨迹移动特征

Tab. 1 Customer trajectory's movement feature

特征名称	特征计算方式
顾客室内停留时间	$LP_{i+1}[time] - LP_i[time]$
顾客室内移动总距离	$\sum_{i=0}^{n-1} Dist(LP_i, LP_{i+1})$
顾客室内平均移动速度	$\sum_{i=0}^{n-1} Dist(LP_i, LP_{i+1}) / (LP_{i+1}[time] - LP_i[time])$
顾客各楼层停留时间	各楼层第一个定位点和最后一个时间差

注:表中各个变量含义同2.2节所述。

其中顾客各楼层停留时间采用 one-hot 方法进行表示,以由与楼层数量相同的特征维数记录顾客在每个楼层停留时间。在提取上述轨迹信息后,通过标准化方法对各个变量进行处理,使特征在表示轨迹时不会因为极高值的出现产生难以衡量轨迹相似性的情况。公式为:

$$R = \frac{(X - \bar{X})}{\sigma^2} \quad (5)$$

式中: R 为轨迹移动特征的标准化结果向量; X 为轨迹移动特征原始值向量; \bar{X} 为每维特征的平均值; σ^2 为轨迹移动特征数据方差。

2.3.2 轨迹语义特征提取

为丰富顾客属性信息,使顾客轨迹从仅包含几何位置信息的原始轨迹转换为拥有更丰富文本信息的语义轨迹,研究根据顾客在室内的移动特点,利用 Indoor-STDBSCAN 算法^[24]对顾客原始轨迹的轨迹点进行聚类处理,提取出顾客在室内行动时停留点 SP ,形成代表顾客游览过的店铺的语义轨迹序列 $SemanTraj$ 。

通过统计商场顾客到访过的店铺的差异,就可以实现对顾客消费特征的提取。考虑到商场室内环境下各个店铺的差异主要体现在其市场定位不

同,因此研究加入从网络图商获取的店铺类型信息以及从大众点评和京东获取的店铺平均消费价格信息作为店铺的文本附加信息,本研究最终选择轨迹语义特征如表2所示,通过对每位顾客的语义轨迹进行统计得到对应的轨迹语义特征。

表2 轨迹语义特征

Tab. 2 Customer trajectory's semantic feature

特征名称	特征计算方式
顾客访问店铺数量	$\text{Length } \text{SemanTraj}_{\max}$
顾客在某店铺中停留最长时间	$\text{Max} (\{SP_i[\text{LevT}] - SP_i[\text{ArrT}]\}_{i=0}^n)$
顾客游览店铺总时间	$\sum_{i=0}^n (SP_i[\text{LevT}] - SP_i[\text{ArrT}])$
顾客消费水平平均值	$\sum_{i=0}^n (SP_i[\text{Price}]) / n$
顾客消费水平最大值	$\text{Max} (\{SP_i[\text{Price}]\}_{i=0}^n)$
顾客消费能力标准差	$\text{Std} (\{SP_i[\text{Price}]\}_{i=0}^n)$
顾客访问的各种类型的店铺数量	语义轨迹中包含的各种类型店铺数量

注:表中各个变量含义同2.2节所述。

上述信息分别从顾客移动特征、消费水平、对不同类型店铺的喜好视角对顾客进行了表示。顾客访问各店铺类型数量同样采用 one-hot 的方法进行表示。提取以上特征后,通过标准化方法对以上数据在每个维度上进行标准化处理,公式为:

$$S = \frac{(T - \bar{T})}{\sigma^2} \quad (6)$$

式中: S 为轨迹语义特征的标准化结果特征向量; T 为轨迹语义特征原始值向量; \bar{T} 为数据均值; σ^2 为轨迹语义特征数据方差。

2.3.3 顾客嵌入特征提取

由于室内同类型店铺的功能定位存在较高相似性,因此在大量顾客轨迹中会出现相似店铺多次同时出现在不同顾客语义轨迹中的情况,通过这种共现关系各个店铺间的属性、空间相似关系就可以作为规律从大量顾客轨迹中提取得到。研究提出利用顾客语义轨迹中各个店铺间的共现情况,通过表示学习 GloVe 算法^[30],训练得到每个店铺的嵌入式特征向量表示,并以店铺嵌入特征为基础对顾客进行嵌入式表示。

GloVe 算法利用所有顾客的语义轨迹数据集中的各个店铺的共现频率构建的共现矩阵进行店铺嵌入向量的学习。考虑到室内空间中对各区域分工设计明确的特点,顾客对店铺偏好会从始至终表现在轨迹中,因此在构建 GloVe 算法的共现矩阵时

不设置统计窗口,而是对完整语义轨迹中店铺共现情况进行统计,通过室内顾客轨迹数据集统计构建 GloVe 共现矩阵算法如下:

算法1 Indoor-GloVe 共现矩阵构建算法

```

输入: 轨迹数据集:  $\text{SemanTrajSet} = \{\text{SemanTraj}_i\}_{i=0}^{\text{NumUsers}}$ 
输出: 店铺共现矩阵: CoocurMatrix
Function: Cooccurrence_Statistic ( $\text{SemanTrajSet}$ ):
1 初始化共现频率矩阵:  $\text{CoocurMatrix} = [0]_{N \times N}$ 
2 For next  $\text{SemanTraj}_i \in \text{SemanTrajSet}$  do
3   For next  $SP_1 \in \{\text{SemanTraj}_i\}_{i=0}^{n-1}$  do
4     For next  $SP_2 \in \{\text{SemanTraj}_i\}_{i=SP_1+1}^n$  do
5        $\text{CoocurMatrix} [SP_1, SP_2] += 1$ 
6 return CoocurMatrix

```

GloVe 算法的损失函数如式(7)、式(8)所示。

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \quad (7)$$

$$f(x) = \begin{cases} (x/x_{\max})^{0.75} & x < x_{\max} \\ 1 & \text{其他} \end{cases} \quad (8)$$

式中: w_j 是店铺 j 的向量; b_i 是店铺 i 的偏差项; X_{ij} 为共现矩阵在 i 行 j 列处的值; V 为店铺的数量,即共现矩阵尺寸。通过输入店铺共现矩阵,对店铺向量进行基于梯度下降的训练,得到各个店铺的店铺特征向量。与其他表示学习方法相比,GloVe 算法充分利用了历史轨迹中的全局统计信息,在室内空间店铺间空间、属性联系密切的情况下,可有效提取店铺的特征信息,对店铺进行嵌入式表示。

具有不同移动性的顾客进入室内环境时倾向于经过的语义点有所区别,在获取到每个店铺的特征向量后,就可以通过顾客的语义轨迹对顾客进行特征建模。不同的特征集成方法侧重集成的特征会有较大差异,研究采用了求取顾客语义轨迹中店铺特征向量的平均值的方法,构建顾客特征向量,由于顾客语义轨迹长度变化较大,因此相比于求和的方法,平均值可以更好地反映顾客对于店铺的偏好。

$$U = \sum_{i=1}^n S_k / n \quad (9)$$

式中: U 为顾客移动特征向量; S_k 为 K 维的店铺向量,取自该顾客语义轨迹中的所有店铺; n 为该顾客语义轨迹长度,即经过的店铺数量。

2.3.4 顾客消费特征组合

本研究提取了室内商场顾客在轨迹移动特征、消费行为特征、移动特征3个方面的特征表示,通过特征向量拼接的方法(Concatenate),将每个包含多维向量的特征进行组合,形成每位商场顾客具有多

源特征的消费特征向量,使在消费模式分析过程中,多视角特征都可以得到考虑。

$$\text{ConFeature} = \text{concatenate}(R, S, U) \quad (10)$$

式中: ConFeature 代表每位商场顾客的顾客特征向量; R 代表原始轨迹中提取的轨迹移动特征向量; S 代表轨迹语义特征向量; U 代表顾客嵌入特征向量。其中向量 ConFeature 的特征维数为 R 、 S 、 U 三个特征向量的维数之和。

2.3.5 顾客消费行为聚类推断

由于商场顾客自身消费属性和室内店铺的相似性,室内环境存在大量消费特征相似的顾客群体,根据生成的顾客消费特征向量通过聚类算法提取出不同的顾客群体,再通过分析同聚类内的消费特征相似性和聚类间特征的差异,实现对室内消费者进行消费行为分析推断。

考虑到多源特征组合产生的向量维度较高,本研究使用了自组织神经网络(Self-Organized Map, SOM)^[31]和谱聚类(Spectral Clustering)^[32] 2种算法以余弦相似度作为距离度量进行聚类。聚类算法需要通过样本间的相似性来确定样本关系,而在高维样本空间下通过欧氏距离进行距离度量存在距离值趋于相同,相似性难以比较的问题,而余弦相似度可以在一定程度上克服高维距离消失的现象,研究引入的聚类算法通过构建神经网络或图的结构,只需要根据样本间的距离关系就可以实现聚类的划分,可以更好地适应余弦相似度(式(11))。

$$\text{cosine_similarity}(A, B) = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (11)$$

式中: A, B 为 n 维向量; A_i, B_i 为其中的一维特征。SOM算法通过网络竞争原则,以网络权重与数据特征间的余弦相似度为度量,更新优胜神经元以及邻近神经元的权重,临近神经元权重更新函数 σ 如式(13)所示。

$$\sigma(w_i, w_{\text{winner}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left(\frac{\|w_i - w_{\text{winner}}\|}{\sigma}\right)^2\right) \quad (12)$$

式中: w_i 为临近神经元; w_{winner} 为优胜神经元; $\|w_i, w_{\text{winner}}\|$ 范数表示二者网络邻接距离; σ 为距离衰减系数,研究中取为2。通过逐次更新优胜神经元以及邻近神经元权重,实现整个网络对高维拓扑形状的逼近。具体算法如算法2所示。

算法2 顾客消费特征SOM处理算法

输入:顾客消费特征数据集: $\text{ConFeatureSet} = \{\text{ConFeature}_i\}_{i=0}^{\text{NumUsers}}$
 神经元边长数: N
 神经元特征数: n
 迭代阈值: num_iteration
 学习率: λ
 输出:神经元权重矩阵: WeightMatrix
 Function: $\text{SOM_ConFeature}(\text{ConFeatureSet}, N, \text{num_iteration}, \lambda, n)$:
 1 随机初始化神经元权重矩阵: $\text{WeightMatrix} = [[w_{ij}]_{N \times n}$
 2 For num_iteration do
 3 For next $\text{ConFeature}_i \in \text{ConFeatureSet}$ do
 4 $\text{winner} = \arg \max \text{cosine_similarity}(w_j, \text{ConFeature}_i)$
 5 $\Delta w_{\text{winner}} = \lambda(\text{ConFeature}_i - w_{\text{winner}})$
 6 $\Delta w_i = \lambda \cdot \sigma(w_i, w_{\text{winner}})(\text{ConFeature}_i - w_{\text{winner}})$
 7 return WeightMatrix

SOM可以提取高维空间中数据的拓扑信息,即相似数据的邻近关系,通过可视化SOM的网络节点与其邻近神经元距离总和图,可直观的体现顾客特征向量间的距离关系,本研究将其作为确定商场顾客聚类数量的依据。

通过SOM降维图观察空间中聚类数量后,研究通过谱聚类确定最终的聚类结果。谱聚类将顾客视为无向图中的各个节点,同样以余弦相似度计算顾客节点间相似性,构建出顾客图的连接权重矩阵 W ,研究以全连接的方法构建无向图的连接矩阵 D ,并通过连接度矩阵计算出顾客相似度图的拉普拉斯矩阵 L ,谱聚类降维原理如式(13)、式(14)所示。

$$L = D - W \quad (13)$$

$$\begin{cases} \arg \min_F \text{tr}\left(F^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} F\right) \\ \text{s.t. } F^T F = I \end{cases} \quad (14)$$

式中: L 为图的拉普拉斯矩阵; D 为图的度矩阵, W 为顾客的相似性矩阵; F 为谱聚类的特征值矩阵。通过在特征值矩阵转置与其自身相乘为单位阵的条件下,通过最小化NCut切图方法的目标函数,以对邻接矩阵的最优分割为优化目标下,提取指定维数的新的特征矩阵 F 。此时的特征矩阵已经被降至低维,且包含有最优聚类分割信息,通过K-Means聚类算法即可实现聚类。由于谱聚类建立在图论基础上,并且以图节点的分割作为降维目标,因此谱聚类在面对大量样本的高维度特征时聚类表现更好,并能收敛于任意形状样本空间的特点,从而在进行顾客特征聚类时表现优于其他聚类方法。

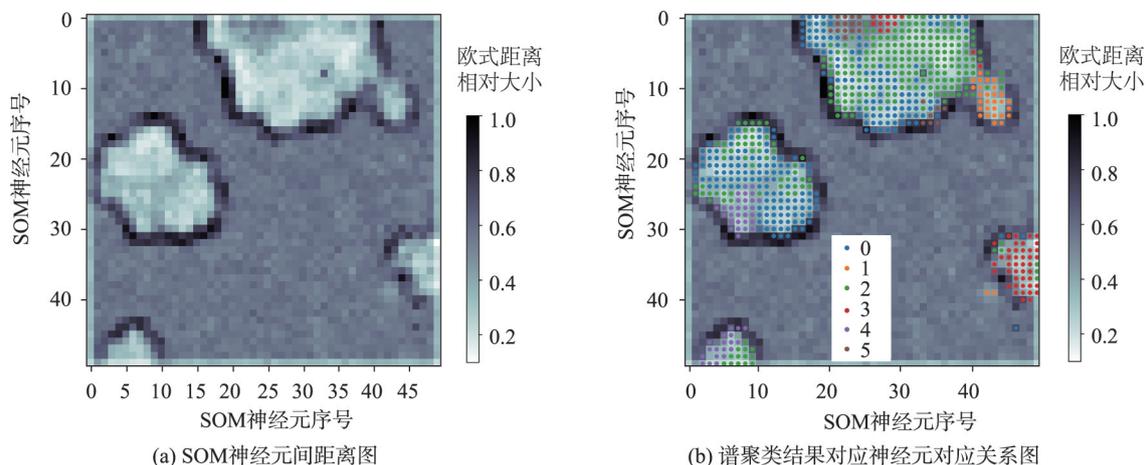


图4 SOM神经元间距离与谱聚类结果对应

Fig. 4 Comparison of SOM's result and spectral clustering's result

相似度较高的一个簇。通过观察图4(a)中神经元间距离,将顾客分为6类,利用谱聚类对顾客特征向量进行聚类,再将每个聚类结果投射到SOM网络中与该顾客特征最相似的神经元,结果如图4(b)所示,不同颜色的点代表谱聚类的聚类结果,共计6类,可见谱聚类结果与SOM神经元识别出的聚类在特征高维空间中的分布基本吻合,故研究对谱聚类识别出的6类顾客进行消费行为模式分析。

根据聚类结果,计算每一类顾客语义轨迹中店铺的消费价格的平均值、平均访问店铺数量、平均移动距离等数据的统计值,如表5所示,各聚类的统计值可作为对每类商场顾客群体消费行为差异的分析依据。

从消费价格平均值的统计值可见,聚类结果可以充分地体现顾客间消费水平的差异,中消费水平的聚类1占大多数,高消费水平聚类2顾客数量较少,低消费水平聚类0占比相比于高消费水平较多。

3.3 顾客消费行为模式分析

顾客在室内购物时,不同类型的商场顾客群体的常访店铺在种类上会有很大的区别,研究通过词云的方法,对每种类型的顾客的常访店铺进行可视化,访问频数决定词大小,并根据不同的店铺类型进行词云颜色渲染,结果如图5所示。

结合词云中的聚类内顾客群体的常访店铺及其类型,结合通过各个聚类内顾客特征的统计值,研究分析得出以下推论:

(1)聚类1作为消费者数量最多的聚类,常访店铺类型以各种服饰类、娱乐类店铺为主,店铺类型混杂,并无显著主要店铺类型,但从空间视角来看该聚类中的常访店铺多位于商场的楼梯、扶梯附近,该类顾客购物目的性较弱,其主要活动模式在很大程度上受到室内环境设计的影响。

(2)从店铺类型来看,聚类0和聚类4的店铺皆以餐饮类店铺为主,饮食类店铺的平均消费价格较低,这就使以餐饮为购物主要目的的顾客在统计消费水平时明显低于其他消费群体,并且该类顾客访

表5 各顾客聚类关键特征统计

Tab. 5 Statistical features of different consumers clusters

顾客类别	顾客数量/人	消费价格均值/元	平均访问店铺数量/个	平均移动距离/m
0	1193	561.1	7.3	2174.7
1	5341	728.4	10.4	2200.5
2	284	1180.3	10.7	2339.6
3	229	746.0	11.3	2603.4
4	470	705.6	6.0	2063.8
5	328	651.6	10.4	2878.8

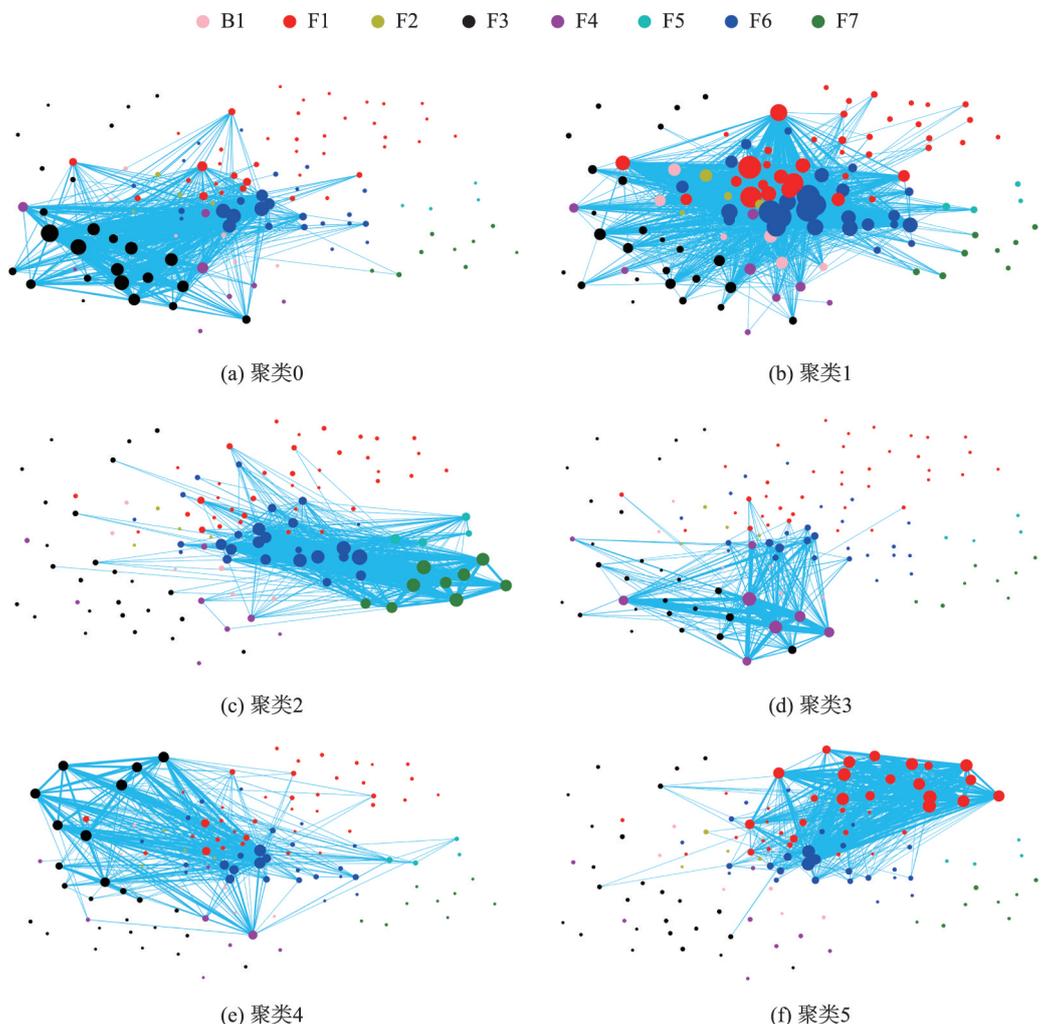


图6 各顾客聚类内人群在店铺间流动关系

Fig. 6 Movement flow of consumers in different clusters between stores

样密切相关,大多数商场在设计时以楼层为单位设置不同功能区,顾客的某一购物目的只需要访问单一楼层即可达到,因此表现出来就是有明确购物目的顾客只要访问一层就能完成购物离开商场。

可见室内环境下的顾客移动行为,除了直接受到顾客自身消费能力、属性特征、消费目的等方面的影响,还在很大程度上取决于室内环境的设计。并且室内环境中,垂直空间上的分异产生的顾客移动代价,往往远大于水平环境,这也就导致了顾客移动规律在垂直层面上的表现出的差异更为明显。

3.5 方法评价

研究使用轮廓系数(Silhouette Coefficient)^[35]进行聚类效果的评价,轮廓系数公式如式(15)所示。

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (15)$$

式中: $a(i)$ 为样本 i 到同簇样本的平均余弦相似度; $b(i)$ 为样本 i 到其他簇的平均余弦相似度。轮廓系数的取值范围为 $[-1, 1]$, 在该区间内值越接近 1 说明样本聚类越合理, 即类内更相似, 类间更分离, 通过求取所有样本轮廓系数的平均值对聚类效果进行评价。

通过对比不同消费行为建模方法在不同聚类算法下的轮廓系数, 评价消费行为的建模方法产生特征的区分度, 以及消费行为聚类划分的显著性。参与比较的聚类算法包括, 研究选用的基于余弦相似度的谱聚类方法、K-Means 算法^[36]、层次聚类算法^[37]; 消费行为建模方法则包括, 研究提出的多源特征组合法、包含嵌入特征方法、包含轨迹语义特征和轨迹移动特征的方法。比较结果如表 6 所示。

表6 特征组成-聚类方法轮廓系数对比

Tab. 6 Silhouette Coefficient comparison of different cluster algorithms and user feature combination

消费行为建模方法	轮廓系数		
	研究方法	K-Means	层次聚类
轨迹移动特征+轨迹语义特征	0.2395	-0.0159	0.1180
顾客嵌入特征	0.2728	0.3028	0.2911
多源特征组合	0.4067	0.2335	0.3868

通过表6中轮廓系数的对比发现,本文提出的通过多源特征组合方法和基于余弦相似度的谱聚类算法结合进行消费行为聚类产生结果的轮廓系数最高。而以多源特征组合作为顾客特征表示产生的聚类结果的轮廓系数也是普遍最高的,相比于单一特征,研究使用的聚类方法对轮廓系数的提升最高达69.8%。可见通过结合网络文本信息和顾客嵌入特征进行商场顾客室内消费特征提取,提取特征能更有效区分室内的消费行为。

4 讨论

本文提出的消费行为分析框架无需隐私敏感的顾客消费记录,依靠轨迹和文本进行顾客消费行为推断,在无法获取商场顾客消费信息的情况下,依靠判断顾客空间上进入店铺的行为形成的语义轨迹序列进行消费行为推断。虽然无法得知进入店铺的顾客是否发生真实的消费行为的情况,会给对顾客消费行为推断造成一定影响,但是从顾客群体角度来看,具有不同消费水平的人群在访问店铺时构成的消费行为会规律性的体现在轨迹中,这样的规律通过历史轨迹的积累就可以体现在研究提取的嵌入式特征中。因此在个体语义轨迹长度较长、群体历史轨迹数量充足的情况下,通过轨迹语义特征可以实现对顾客消费属性的初步划分,再通过加入顾客嵌入特征,就能实现结合群体移动规律的消费特征提取,从而更有效的划分顾客消费水平、提取消费倾向。

在通过表示学习生成嵌入式向量时,由于嵌入特征训练时的初始权重为随机生成,即使使用相同轨迹数据集,在进行多次训练时,最终生成的特征向量也会在数值上不尽相同,但是各店铺嵌入向量间的相似关系基本不变。实验表明,初始权重不同给顾客消费行为聚类分析带来影响极小,

各聚类顾客数量和各聚类内消费水平统计量基本保持不变。

由于室内空间结构复杂,顾客移动行为所受影响因素较多,因此移动规律往往蕴含于完整的轨迹中,这就使共现矩阵这样的全局规律进行学习的GloVe算法,在室内店铺嵌入特征学习中的表现远优于以单个语义轨迹中局部信息为学习目标的嵌入向量生成方法Word2Vec^[38],以及矩阵分解(Matrix Factorization)^[39]这样的以直接的顾客访问矩阵为输入,未经过全局特征统计的嵌入特征提取方法。

5 结论

本文提出了一种基于多视角的顾客室内消费特征轨迹提取方法,方法从轨迹移动特征、轨迹语义特征、顾客嵌入特征3个视角出发刻画顾客的消费行为,并通过提取的消费行为特征对轨迹进行聚类分析,推断不同顾客群体在室内空间的消费特征。真实室内轨迹实验分析表明,大型商场顾客具有明显的消费水平分层,并提取出了对不同类型店铺有明显偏好的顾客群体。本研究提出方法可以普适的刻画大量室内顾客的消费行为特征,通过引入表示学习的方法提取轨迹嵌入特征,实现消费行为中的移动规律学习与消费时空行为模式提取,有效挖掘相似消费行为顾客与不同类型的顾客在室内消费行为特征,为精细尺度人类移动模式研究提供支持。

研究通过将顾客消费行为推断方法应用于真实室内轨迹数据集,发现在商场室内空间中,位于扶梯口、出入口等顾客必经之路上的店铺在各类消费群体中都有较高访问量,并且有明确购物目的的顾客,在室内活动时明显的楼层倾向性。结合室内环境设计的特点,研究认为顾客在大型商场室内空间的消费行为会受到扶梯位置、店铺位置、功能区划分等室内环境设计因素的显著影响。因此在不同商场的室内环境下顾客消费行为也会表现出不同的规律,未来工作可以通过对比更多室内购物环境下顾客消费行为,实现提取室内购物空间顾客消费行为的普适规律。同时也可以在本研究提出的特征提取框架上继续拓展,加入诸如店铺评论等的更加丰富的店铺语义信息,辅助消费行为推断。

参考文献(References):

- [1] Barari M, Ross M, Surachartkumtonkun J. Negative and positive customer shopping experience in an online context[J]. *Journal of Retailing and Consumer Services*, 2020, 53:101985. DOI:10.1016/j.jretconser.2019.101985
- [2] 傅辰昊,周素红,闫小培,等.广州市零售商业中心的居民消费时空行为及其机制[J].*地理学报*,2017,72(4):603-617. [Fu C H, Zhou S H, Yan X P, et al. Spatio-temporal characteristics and influencing factors of consumer behavior in retailing centers: A case study of Guangzhou in Guangdong Province[J]. *Acta Geographica Sinica*, 2017, 72(4):603-617.] DOI:10.11821/dlxb201704004
- [3] 柴彦威,王茂军.日本消费者行为地理学研究进展[J].*地理学报*,2004,59(z1):167-174. [Chai Y W, Wang M J. Progress of geographical study on consumer behavior in Japan[J]. *Acta Geographica Sinica*, 2004,59(z1):167-174.] DOI:10.3321/j.issn:0375-5444.2004.z1.023
- [4] 吴康敏,王洋,叶玉瑶,等.广州市零售业态空间分异影响因素识别与驱动力研究[J].*地球信息科学学报*,2020,22(6):1228-1239. [Wu K M, Wang Y, Ye YY, et al. A study on the influencing factors and driving forces of spatial differentiation of retail formats in Guangzhou[J]. *Journal of Geo-information Science*, 2020,22(6):1228-1239.] DOI:10.12082/dqxxkx.2020.190353
- [5] 张文佳,柴彦威.居住空间对家庭购物出行决策的影响[J].*地理科学进展*,2009,28(3):362-369. [Zhang W J, Chai Y W. The influence of residential space on household shopping tour decision-making behaviors[J]. *Progress in Geography*, 2009,28(3):362-369.] DOI:10.11820/dlkxjz.2009.03.006
- [6] 柴彦威,翁桂兰,沈洁.基于居民购物消费行为的上海城市商业空间结构研究[J].*地理研究*,2008,27(4):897-906. [Chai Y W, Weng G L, Shen J. A study on commercial structure of Shanghai Based on residents' shopping behavior[J]. *Geographical Research*, 2008,27(4):897-906.] DOI:10.3321/j.issn:1000-0585.2008.04.018
- [7] 刘学,甄峰,张敏,等.网上购物对个人出行与城市零售空间影响的研究进展及启示[J].*地理科学进展*,2015,34(1):48-54. [Liu X, Zhen F, Zhang M, et al. Research review of online shopping impact on personal travel and urban retail space and implications[J]. *Progress in Geography*, 2015,34(1):48-54.] DOI:10.11820/dlkxjz.2015.01.006
- [8] 张文忠,李业锦.北京城市居民消费区位偏好与决策行为分析——以西城区和海淀中心地区为例[J].*地理学报*, 2006,61(10):1037-1045. [Zhang W Z, Li Y J. Study on residents' shopping location preference and shopping location choice: A case on Xicheng and Haidian districts of Beijing[J]. *Acta Geographica Sinica*, 2006,61(10):1037-1045.] DOI:10.3321/j.issn:0375-5444.2006.10.003
- [9] 周素红,林耿,闫小培.广州市消费者行为与商业业态空间及居住空间分析[J].*地理学报*,2008,63(4):395-404. [Zhou S H, Lin G, Yan X P. The relationship among consumer's travel behavior, urban commercial and residential spatial structure in Guangzhou, China[J]. *Acta Geographica Sinica*, 2008,63(4):395-404.] DOI:10.3321/j.issn:0375-5444.2008.04.009
- [10] 杨洁,杨乃,黄婷,等.大型商场内人群择路行为认知规律的研究[J].*武汉大学学报·信息科学版*,2017,42(3):414-420. [Yang J, Yang N, Huang T, et al. Cognitive rules of people choosing routes in large stores[J]. *Geomatics and Information Science of Wuhan University*, 2017,42(3):414-420.] DOI:10.13203/j.whugis20140573
- [11] Su D N, Duong T H, Dinh M, et al. Behavior towards shopping at retailers practicing sustainable grocery packaging: The influences of intra-personal and retailer-based contextual factors[J]. *Journal of Cleaner Production*, 2021,279(10):123683.1-123683.12. DOI:10.1016/j.jclepro.2020.123683
- [12] 李也.张量聚类和回归建模及其在消费行为分析上的应用研究[D].上海:上海交通大学,2020. [Li Y. Tensor clustering and regression modeling and its application research in consumer behavior analysis[D]. Shanghai: Shanghai Jiao Tong University, 2020.] DOI:10.27307/d.cnki.gsjtu.2020.002768
- [13] Arnold M J, Reynolds K E. Affect and retail shopping behavior: Understanding the role of mood regulation and regulatory focus[J]. *Journal of Retailing*, 2009,85(3):308-320. DOI:10.1016/j.jretai.2009.05.004
- [14] Papouškova M, Hajek P. Two-stage consumer credit risk modelling using heterogeneous ensemble learning[J]. *Decision support systems*, 2019,118:33-45. DOI:10.1016/j.dss.2019.01.002
- [15] 陈锐志,郭光毅,叶锋,等.智能手机音频信号与MEMS传感器的紧耦合室内定位方法[J].*测绘学报*,2021,50(2):143-152. [Chen R Z, Guo G Y, Ye F, et al. Tightly-coupled integration of acoustic signal and MEMS sensors on smartphones for indoor positioning[J]. *Acta Geodaetica et Cartographica Sinica*, 2021,50(2):143-152.] DOI:10.11947/j.AGCS.2021.20200551
- [16] 陈锐志,王磊,李德仁,等.导航与遥感技术融合综述[J].*测绘学报*,2019,48(12):1507-1522. [Chen R Z, Wang L, Li D R, et al. A survey on the fusion of the navigation and the remote sensing techniques[J]. *Acta Geodaetica et Cartographica Sinica*, 2019,48(12):1507-1522.] DOI:10.11947/j.AGCS.2019.20190446
- [17] Trivedi A, Silverstein K, Strubell E, et al. WiFiMod:

- Transformer-based indoor human mobility modeling using passive sensing[C]. ACM SIGCAS Conference on Computing and Sustainable Societies, 2021:126-137. DOI:10.1145/3460112.3471951
- [18] Hosseini K S, Azaddel M H, Nourian M A, et al. Improving multi-floor WiFi-based indoor positioning systems by fingerprint grouping[C]. the 5th International Conference on Internet of Things and Applications, 2021:1-6. DOI: 10.1109/IoT52625.2021.9469602
- [19] Oosterlinck D, Benoit D F, Baecke P, et al. Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits[J]. Applied Geography, 2017, 78:55-65. DOI:10.1016/j.apgeog.2016.11.005
- [20] Fu X, Zhang H, Wang P. Automatic construction of indoor 3D navigation graph from crowdsourcing trajectories [J]. ISPRS International Journal of Geo- information, 2021,10(3):146. DOI:10.3390/ijgi10030146
- [21] Vanclooster A, Vanhaeren N, Viaene P, et al. Turn calculations for the indoor application of the fewest turns path algorithm[J]. International Journal of Geographical Information Science, 2019,33(11):2284-2304. DOI:10.1080/13658816.2019.1630629
- [22] Wang P, Gao F, Zhao Y, et al. Detection of indoor high-density crowds via Wi-Fi tracking data[J]. Sensors, 2020, 20(18):5078. DOI:10.3390/s20185078
- [23] Mou N, Wang H, Zhang H, et al. Association rule mining method based on the similarity metric of tuple-relation in indoor environment[J]. IEEE Access, 2020,8:52041-52051. DOI:10.1109/ACCESS.2020.2980952
- [24] Wang P, Wang H, Zhang H, et al. A hybrid Markov and LSTM model for indoor location prediction[J]. IEEE Access, 2019,7:185928-185940. DOI:10.1109/ACCESS.2019.2961559
- [25] Zhu J, Cheng D, Zhang W, et al. A new approach to measuring the similarity of indoor semantic trajectories[J]. ISPRS International Journal of Geo- information, 2021,10(2):90. DOI:10.3390/ijgi10020090
- [26] Pei T, Liu Y, Shu H, et al. What influences customer flows in shopping malls: Perspective from indoor positioning data[J]. ISPRS International Journal of Geo-Information, 2020,9(11):629. DOI:10.3390/ijgi9110629
- [27] 于邓波.基于室内行人定位轨迹的行为模式识别与分析[D].武汉:武汉大学,2019. [Yu D B. Pedestrian behavior pattern recognition and analysis of indoor location data [D]. Wuhan: Wuhan University, 2019.] DOI:10.27379/d.cnki.gwhdu.2019.000032
- [28] Liu Y, Cheng D, Pei T, et al. Inferring gender and age of customers in shopping malls via indoor positioning data [J]. Environment and Planning B: Urban Analytics and City Science, 2020,47(9):1672-1689. DOI:10.1177/2399808319841910
- [29] Yaeli A, Bak P, Feigenblat G, et al. Understanding customer behavior using indoor location analysis and visualization[J]. IBM Journal of Research and Development, 2014, 58(5/6):3:1-3:12. DOI:10.1147/jrd.2014.2337552
- [30] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing, 2014:1532-1543. DOI:10.3115/v1/D14-1162
- [31] Kohonen T. The self-organizing map[J]. Proceedings of the IEEE, 1990,78(9):1464-1480. DOI:10.1016/S09252312(98)00030-7
- [32] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]. Proceedings of the Advances in neural information processing systems, 2002:849-856. DOI:10.1.1.19.8100
- [33] Guo X, Ansari N, Hu F, et al. A survey on fusion-based indoor positioning[J]. IEEE Communications Surveys & Tutorials, 2019,22(1):566-594. DOI:10.1109/COMST.2019.2951036
- [34] Kobourov S G. Spring embedders and force directed graph drawing algorithms[J]. CoRR, 2012,abs/1201.3011. arXiv:1201.3011
- [35] Peter R J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics, 1987,20:53-65. DOI:10.1016/0377-0427(87)90125-7
- [36] J A Hartigan, M A Wong. A K-Means clustering algorithm [J]. Journal of the Royal Statistical Society, 1979,28(1):100-108. DOI:10.2307/2346830
- [37] D'Andrade R G. Hierarchical clustering[J]. Psychometrika, 2011,43(1):59-67. DOI:10.1007/BF02294089
- [38] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. CoRR, 2013,abs/1301.3781. arXiv:1301.3781
- [39] Takács G, Pilászy I, Németh B. Matrix factorization and neighbor based algorithms for the Netflix prize problem [C]. Proceedings of the 2008 ACM conference on Recommender systems, 2008:267-274. DOI:10.1145/1454008.1454049