计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(9):32−39 [doi: 10.15888/j.cnki.csa.007600]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

空间位置的关联分析及其向量化表示方法①

张　舒[1,2], 郭旦怀[1,2], 周纯葆[1,2], 李薰春[3], 靳　薇[4,5]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](国家广播电视总局广播电视科学研究院, 北京 100866)
[4](北京市科学技术研究院, 北京 100089)
[5](北京市新技术应用研究所, 北京 100094)
通讯作者: 郭旦怀, E-mail: guodanhuai@cnic.cn

摘　要: 理解地理空间位置的空间相关性, 对于地理信息检索、推荐系统, 城市交通管理, 居民出行模式探究等应用研究具有重要支撑作用. 为更具体表义空间位置及其关联关系, 本文基于多种居民出行轨迹数据, 提出一种基于深度学习的空间位置向量化表示方法, 而后通过空间位置向量的向量运算, 可计算得到空间位置的关联程度. 首先将长、短距离出行轨迹进行匹配连接, 构建大规模交通网络, 覆盖多种出行模式, 得到对不同位置间空间关联信息的完整识别. 然后基于图神经网络模型, 本文提出融合位置特征与轨迹信息的空间向量化表示方法, 并优化其训练学习中节点采样方法, 提高空间向量的表达能力. 最后以北京市共享单车轨迹数据与公共交通路网数据进行实证分析, 实验结果表明基于本文提出方法生成的空间向量在空间位置的关联分析、聚类分析中相比 DeepMove 等已有方法拥有更好的效果.

关键词: 空间关联分析; 空间向量; 图神经网络; 轨迹数据

引用格式: 张舒,郭旦怀,周纯葆,李薰春,靳薇.空间位置的关联分析及其向量化表示方法.计算机系统应用,2020,29(9):32–39. http://www.c-s-a.org.cn/1003-3254/7600.html

Correlation Analysis and Vectorization Method for Spatial Position

ZHANG Shu[1,2], GUO Dan-Huai[1,2], ZHOU Chun-Bao[1,2], LI Xun-Chun[3], JIN Wei[4,5]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Academy of Broadcasting Science, National Radio and Television Administration, Beijing 100866, China)
[4](Beijing Academy of Science and Technology, Beijing 100089, China)
[5](Beijing Institute of New Technology Application, Beijing 100094, China)

Abstract: Understanding the spatial correlation of places plays an important role in geographic information retrieval and recommendation systems, urban traffic management, and resident travel pattern exploration. In order to represent the places and their spatial relationships specifically, we propose a deep learning-based vectorization method for places. The correlation between places can be calculated by the place vectors. Firstly, the trajectories of long-distance and short-distance are matched and connected to build a large-scale traffic network, which could cover multiple travel modes and obtain a complete cognition of spatial relations. Then we propose a spatial vectorization method which is based on graph neural network and combines place features and trajectory information. Besides, we improve the representation ability of latent representations for places by optimizing a node sampling method. Finally, the empirical analysis is performed on the shared bicycle track data and public traffic data in Beijing. The result demonstrates that the proposed method outperforms

① 基金项目: 国家自然科学基金 (41971366, 91846301); 国家重点研发计划 (2018YFC0809700); 北京市自然科学基金 (9172023, 9194027)
Foundation item: National Natural Science Foundation of China (41971366, 91846301); National Key Research and Development Program of China (2018YFC0809700); Natural Science Foundation of Beijing Municipality (9172023, 9194027)
收稿时间: 2020-02-27; 修改时间: 2020-03-17; 采用时间: 2020-03-24; csa 在线出版时间: 2020-09-04

the existing methods such as DeepMove on place correlation analysis and cluster analysis.

Key words: spatial correlation analysis; spatial representation; graph neural network; trajectory data

空间位置作为重要的空间特征, 常用于目的地预测与推荐, 城市功能规划, 交通管控系统, 交通流量预测、位置分类等应用. 传统的空间位置表示方法与空间向量化表示方法, 常将其经纬度坐标与兴趣点 (Points Of Interest, POI) 信息将空间位置映射为 ID 类型, 其中经纬度用来精确地描述空间位置, 兴趣点为空间位置增添了大量属性特征. 但是, 传统的空间位置表示方法只保留了自身经纬度与 POI 的信息, 缺少了空间位置间隐含的时空关联信息, 难以理解空间位置和推断空间位置间的关联关系.

融合空间位置的时空关联信息与其周边 POI 信息, 将有助于提高空间向量的表达效果. 居民出行轨迹数据能够为结合两种信息的空间分析提供可靠的数据支持. 居民的出行轨迹可体现居民的出行模式与活动规律, 同时也反映出空间位置之间的关联, 是制定城市交通管理方案的关键. 随着 GPS 定位技术的快速发展, 可收集到大规模轨迹数据, 使得轨迹数据能够被用于更加细粒度地分析空间位置的时空关联关系. 当今的大规模轨迹数据, 包括机动车轨迹与非机动车轨迹, 已能够支持机器学习、深度学习等对于数据量的需求, 并且能够为空间位置及其关联关系的理解、表示与推理学习提供可能性.

另一方面, 随着深度学习在自然语言处理、计算机视觉等领域的迅猛发展, 研究者已在空间分析中引入了深度学习方法[1]. 近些年提出的空间上下文理解模型中, 考虑空间位置与其临近位置之间的关系, 结合自然语言处理中的词向量模型, 使用神经网络将词生成固定长度的向量表示, 该词向量表示能包含该词的语义与上下文信息, 最终生成结合空间邻接关系的空间位置向量化表示[2,3]. 但这些方法仅关注了空间的静态特征, 尚未考虑到居民在空间之间的轨迹移动, 从而缺失了居民出行的时空模式关联信息.

在使用深度学习技术处理复杂的关联关系时, 一个有效的模型是图神经网络[4]. 图神经网络因其能够处理图结构形式的数据引起广大关注[5]. 借助图神经网络算法能够为更加复杂的图结构中的节点生成低维向量表示, 该向量表示既包含了节点的类别特征, 又聚合了节点在复杂图网络中的邻域特征[6]. 交通网络以图结构形式存在, 因此, 交通网络常使用图神经网络解决流量预测问题[7,8] 与出发地、目的地预测[9] 等问题. 围绕空间向量化表示任务, DeepMove 基于出租车轨迹数据, 使用图神经网络中的随机游走方法与 Word2Vec 方法, 为交通网络中的 POI 节点生成含有邻域关联信息的向量化表示[10]. 但是 DeepMove 仅使用了简单的图神经网络, 未使用节点自身特征, 难以聚合邻域节点特征. 另一方面, 这些研究只基于单一的轨迹数据, 而没有考虑不同出行方式之间的联系. 例如, 居民从住所前往公司, 通常会产生从住处到地铁站的骑行轨迹、从地铁站到公司所在地的地铁轨迹、公交站点之间的公交轨迹, 通过综合这些轨迹才能发现住处和公司所在地之间存在的轨迹关联, 而对不同类型的轨迹分别处理则会遗漏大量信息.

本文提出基于图神经网络的空间位置向量化表示方法, 综合公交线路轨迹、地铁线路轨迹与大规模共享单车轨迹数据构建多源交通网络. 同时, 针对不同源交通轨迹数据的集成, 本文提出一种针对不同类型轨迹数据的集成方法: 将长距离出行与短距离出行进行匹配连接, 以更全面地覆盖不同位置之间的空间关联. 使用本文方法生成的空间位置的向量化表示, 能够综合空间特征、邻域特征与时空关联特征. 相较于其他已有方法, 本文提出的空间向量化表示方法能够学习到空间位置的关联关系.

本文首先围绕着不同类型轨迹数据的网络构建展开讨论; 其次介绍融合 POI 与轨迹信息的空间向量化表示方法; 然后设计实验以验证本文提出的空间向量化表示方法的有效性; 最后讨论针对空间位置向量化表示任务, 有待探索的研究方向.

1 多源交通轨迹数据网络构建

1.1 多源交通轨迹数据

相较于公交车与地铁等交通方式, 近年来共享单车的流行为人们的出行方法提供了新的选择. 共享单车因其更加灵活与便捷的特点, 使得共享单车轨迹数据相较于其它交通轨迹数据, 具有覆盖范围更广, 数据规模更大, 采样密度更高, 位置精度更高等特点. 由于共享单车在城市交通中的占比逐步增大, 能够从一定

程度上更加细粒度地描述城市居民的短距离出行模式, 因此本文采用公共交通路网数据 (包括公交线路与地铁线路), 与共享单车轨迹数据相结合的方式, 共同构建大规模交通网络. 融合了共享单车轨迹数据与公共交通线路的交通网络能够更完整地体现居民的出行模式.

为了解居民的短距离出行模式, 本文统计了共享单车轨迹数据分布. 居民在工作日与非工作日一天内不同时间段的出行分布图如图 1 所示, 可发现在工作日期间的早晚上班高峰时刻, 居民对于共享单车出行的需求非常大; 周末的出行需求量较为平均. 此外, 在剔除了异常轨迹距离后, 骑行轨迹距离分布图如图 2 所示. 可发现, 当出行距离小于 2 公里时, 居民更倾向选择共享单车出行.

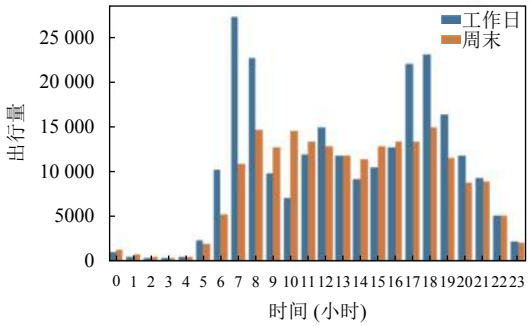


图 1 不同时段内共享单车出行量分布图

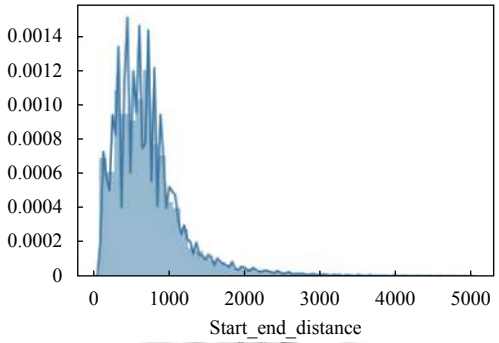


图 2 骑行轨迹距离分布图

对共享单车轨迹数据进行分析后可发现, 利用共享单车轨迹数据与公共交通路网数据相结合的方式构建大规模交通网络, 融合多种交通方式的出行轨迹, 能够覆盖居民针对不同出行需求与出行距离的交通轨迹.

目前通过互联网开源数据, 可获取到大量免费的共享单车轨迹数据以供研究者使用. 本文使用摩拜共享单车开源数据集, 涵盖了北京市居民于 2017 年 5 月 14 日至 2017 年 5 月 21 日的骑行轨迹, 样本量共计 3 214 096 条. 为避免异常数据的影响, 剔除了骑行距离大于 5 公里的轨迹. 此外, 通过数据爬取技术, 获取北京市公共交通路网数据, 具体包括公交车线路数据与地铁线路数据

1.2 集成不同类型轨迹数据的网络建模

为综合不同类型的轨迹数据, 识别居民完整的出行模式, 本文将长、短距离出行轨迹进行匹配连接. 对不同类型轨迹数据的匹配方法如下:

第 1 步. 使用 GeoHash[11] 技术将共享单车轨迹数据与公交交通路网数据中的地点经纬度向量成固定长度的字符串. 该字符串由空间位置的经纬度向量得到, 并且其长短可用于划分空间位置的大小. 本文具体使用长度为 7 位的 GeoHash 字符串.

第 2 步. 使用共享单车轨迹中的出发地与目的地作为节点, 轨迹作为边, 标记边的类型为共享单车出行, 居民在此出发地与目的地的出行次数作为边的权重, 以构建大规模交通网络.

第 3 步. 针对公共交通路网数据, 以公共交通站点作为节点, 线路作为边, 依据公共交通的类型作为边的类型, 构建大规模交通网络.

第 4 步. 通过以上步骤, 合计产生 10.5 万个节点. 通过百度地图开放平台, 获取交通网络中的每个节点的 POI 信息, 经过 one-hot 处理后, 作为节点的属性特征; 获取每个节点的交通拥堵信息, 经过离散化处理后, 作为节点的拥堵特征.

图 3 为使用该方法对长、短距离出行轨迹进行匹配的一个示例, 例如从图中 A-B, B-C, C-D 分别为某居民从住处到公交站的骑行轨迹, 公交运行轨迹, 从公交站到公司的骑行轨迹, 通过匹配以上 3 段轨迹可识别出 A 到 D 点的实际关联.

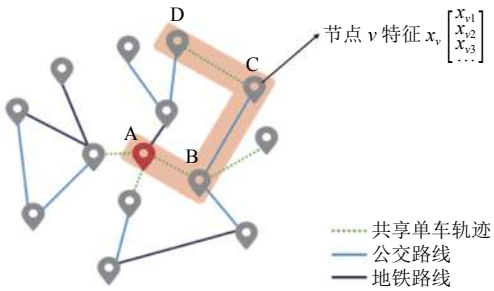


图 3 长、短距离出行轨迹匹配示例

2 融合 POI 与轨迹信息的空间向量化方法

本章节首先给出交通网络的形式化定义; 然后介绍本文提出的 POI 与轨迹信息融合模型; 最后说明基于该模型的空间向量化表示方法.

2.1 形式化定义

文中常见符号的定义如表 1 所示. 使用$G=(V,E)$表示交通网络, $n=|V|$表示节点的个数. 使用 $X\in R^{n\times D}$ 表示节点特征矩阵, x_v表示节点 v 的特征, 其中$x_v\in R^D$, $\forall v\in V$, 表示矩阵 X 的第 v 行第 j 列, 即节点 v 的第 j 个特征. G 与 X 作为图神经网络模型的输入.

表 1 符号定义

符号	含义
G	交通网络
V	空间位置节点集合
E	边集合, 即不同类型的交通轨迹
n	空间位置节点个数
N_v	节点v的邻居节点
$X\in R^{n\times D}$	节点特征矩阵
$x_v\in R^D$	节点v的特征
Y_v	节点v的标签

2.2 POI 与轨迹信息融合建模

为了在神经网络模型中结合空间位置节点的 POI 信息与轨迹关联信息, 本文基于 GraphSAGE 模型[12]提出融合建模的方法. GraphSAGE 模型能够通过对目标节点的邻居进行随机采样得到子图, 再对子图进行卷积, 替代了直接对全图进行卷积的方式, 大大降低了计算和内存的压力. 此外, GraphSAGE 模型通过学习聚合函数 (aggregator) 的方式, 把邻居节点的特征聚合到中心节点自身. 当学习得到聚合函数后, 聚合函数能够泛化到新的节点或者新的网络上, 即使是在训练过程中出现未知的节点, 模型也能推断出其向量化表示. 这一做法替代了直接学习网络节点的向量化表示, 泛化能力更强, 是一种归纳式 (inductive) 学习算法.

基于 GraphSAGE 的向量化表示学习方法在每次迭代中进行以下 3 个步骤:

第 1 步. 对图中的每个节点采样固定数量的邻居节点作为该节点的邻居节点集合;

第 2 步. 通过模型学习的聚合函数 (aggregator) 对采样得到的邻居节点集合进行聚合, 以把邻居集合节点的特征信息聚合到中心节点上, 得到新的节点向量;

第 3 步. 通过聚合邻域特征得到的节点的向量向量化表示用于损失计算, 更新权重矩阵 W.

算法 1 为具体的向量化学习方法. 在算法 1 中, K 为图卷积的层数, 表示每个节点聚合 K 阶邻居. 在外层循环的第 k 次迭代中, 对于每个节点 v 首先通过聚合函数 Agg 来对节点 v 的邻居节点的 k–1 层 embedding 向量进行聚合, 得到节点 v 第 k 层的邻居聚合 embedding, 再将节点 v 的 k–1 层得到的向量拼接起来接入全连接网络层, 最终得到节点 v 在第 k 层的向量化表示.

算法 1. 基于 POI 与轨迹信息融合建模的空间向量化表示算法

输入: 图$G=(V,E)$; 节点特征矩阵 X; 图采样深度 K; 权重矩阵$W^k,\forall k\in K$; 非线性激活函数 f; 聚合函数 Agg; 邻居节点采样函数 S.
输出: 节点的向量化表示$\varphi_v,\forall v\in V$

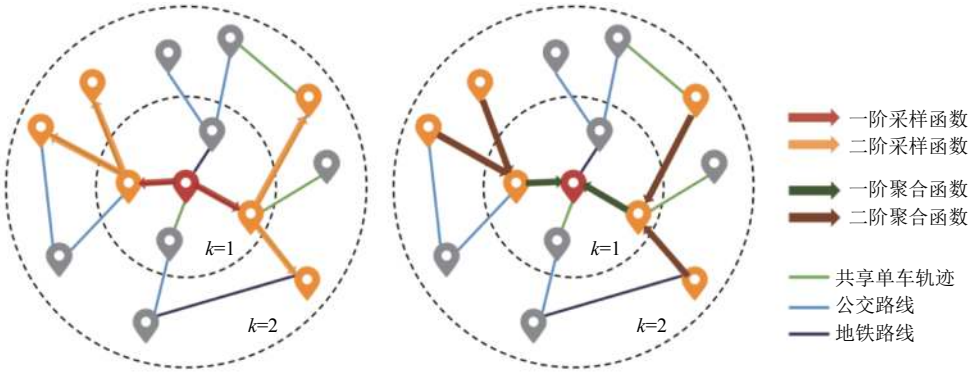
步骤:
$h_v^0\leftarrow x_v,\forall v\in V$
for $k=1\cdots K$ do:
  for $v\in V$ do:
    $h_{S(v)}^k\leftarrow Agg\left(h_i^{(k-1)},\forall i\in S(v)\right)$
    $o_v^k=CONCAT\left(h_v^{(k-1)},h_{S(v)}^k\right)$
    $h_v^k\leftarrow f\left(W^k\cdot o_v^k\right)$
  end
  $h_v^k\leftarrow\frac{h_v^k}{\|h_v^k\|_2},\forall v\in V$
end
$\varphi_v=h_v^K,\forall v\in V$

2.3 空间向量化表示学习方法

已有的 Place2Vec 研究通常使用 POI 的语义特征与空间特征以生成 POI 的向量化表示. 但具体的某个空间位置通常存在多个 POI, 例如在商场与餐厅, 住宅区与超市大多会同时出现, 难以使用单独的 POI 来表示空间位置. 本文提出的方法通过构建大规模多源数据交通网络与 POI 与轨迹信息融合方法直接得到细粒度空间位置的向量化表示. 基于章节 1.2 中使用公共交通路网数据与共享单车轨迹数据构建的大规模交通网络中, 边具有各自的类型与权重, 各节点的邻居数分布不均匀. 因此, 针对空间位置向量化表示任务, 须对图神经网络中的采样方法与聚合方法加以修改, 以适应交通网络的特性. 具体的, 考虑到使用多源数据构建的交通网络的边都具有类别、权重等特征, 并且根据图 2 中对共享单车轨迹数据的统计分析, 可发现居民在短距离出行时更倾向于选择共享单车出行. 因此本文提出一种基于出行距离长短的采样方法, 示意图见图 4(a). 对目标节点 v 的第 k 层邻居采样过程中, 当 k=1 时, 即一阶采样函数, 优先采样边类型为共享单车轨迹的邻居节点, 当 k=2 时, 即二阶采样函数, 优先采样边类型为公共交通线路的邻居节点.

经过采样函数后, 节点的邻居节点集合是无序的, 因此聚合函数不仅需要有很强的表征学习能力, 还具有对称性 (symmetric) 要求, 即函数的输出与输入聚合函数的节点顺序无关. 本文使用文献 [12] 中提出的 GCN aggregator 作为聚合方法, 将不同层级的邻域中的邻居

节点的特征聚合起来, 并且将聚合后的邻域节点特征与目标节点特征拼接结合后, 传递到全连接网络中. 其中节点的聚合过程如图 4(b) 所示. 利用图神经网络的反向传播机制, 最终得到节点的向量化表示.



(a) 邻域节点采样 (b) 邻域节点信息聚合

图 4 交通网络中采样与聚合操作示意图

为了在无监督学习过程中得到更有效的空间位置向量化表示, 使用图神经网络中的无监督损失函数, 该损失函数的优化目标是最大化正样本的概率, 使得邻居节点的向量化表示更加相近; 同时最小化负样本的概率, 使得没有共同交点的节点的向量化表示相异:

$$J_g(z_v) = -\log(f(z_v^{\mathrm{T}} z_i)) - QE_{i_n \sim p_n(i)} \log(f(z_v^{\mathrm{T}} z_{i_n})) \quad (1)$$

式 (1) 中, $z_v (v \in V)$为图神经网络的输出, 即空间位置的向量化表示; i 为与 v 共同出现在一组随机游走上的节点; $i_n \sim p_n(i)$表示 i_n 服从对 i 的负样本采样分布; Q 定义为负样本的个数. 对此, 使用随机梯度下降学习方法与 Mikolov 等[13] 提出的负采样学习优化算法, 用于更新图神经网络的权重矩阵 $W^k (\forall k \in K)$ 与聚合函数中的参数. 当 k=2, 使用 GCN 聚合方法, 网络中节点规模为 10^4 左右时, 本文提出方法的整体参数规模为 10^5 左右. 此外, 由于空间位置包含多个 POI, 可通过空间位置的向量化表示经过加权聚合操作后得到 POI 的向量化表示. 具体的, POI 的向量化表示 p_i 可定义为:

$$p_i = \frac{\sum_{j=1}^{n} w_{ij} \varphi_j}{\sum_{j=1}^{n} w_{ij}} \quad (2)$$

w_{ij} 为位置向量 φ_j在 p_i 中的权重. 当使用平均权重时, w_{ij} = 1. 其中, 本文使用 tf-idf 统计方法计算权重 w_{ij}[14], 用以度量位置与 POI 的相关程度.

3 实验分析

为了验证本文提出的空间位置的向量化表示方法的有效性, 我们分别对空间位置向量化表示、由空间位置向量化表示加权聚合得到的 POI 向量化表示进行评估. 为了直观地理解空间位置的向量化表示, 部分向量聚类结果采用可视化方式展现.

3.1 评价指标

使用空间位置向量化表示间 (v_i, v_j) 的 Cosine 距离[15] 以定义空间位置的相似度 S_{space}:

$$S_{\text{splace}}(v_i, v_j) = \cos(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|} \quad (3)$$

空间位置向量化表示的距离 d_{place} 则为:

$$d_{\text{place}}(v_i, v_j) = 1 - S(v_i, v_j) \quad (4)$$

相似的, 由空间位置向量化表示经过加权聚合操作得到的 POI 向量化表示, 其相似度 S_{poi} 定义为 POI 向量化表示的 Cosine 距离. 此外, 使用轮廓系数评估向量化表示的聚类结果[16]. 轮廓系数 $s(i)$ 为:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

其中, $a(i)$ 表示样本 i 到同簇其他样本的平均距离; $b(i)$ 则为样本 i 到其他簇的最小平均距离; $s(i)$ 的值域在 [−1, +1] 之间. $a(i)$ 度量类内距离, $b(i)$ 度量类间距离. 当 $a(i) << b(i)$ 时, 即类内距离远小于类间距离, 则 $s(i)$ 接近于 1, 表明聚类效果愈好. 反之, 当 $a(i) >> b(i)$ 时, 即类内距离远大于类间距离, 则 $s(i)$ 接近于−1, 表

明聚类效果愈差, 样本 i 更应该被分类到其他簇. 本文使用平均轮廓系数以评估整体样本的聚类结果.

3.2 实验结果与分析

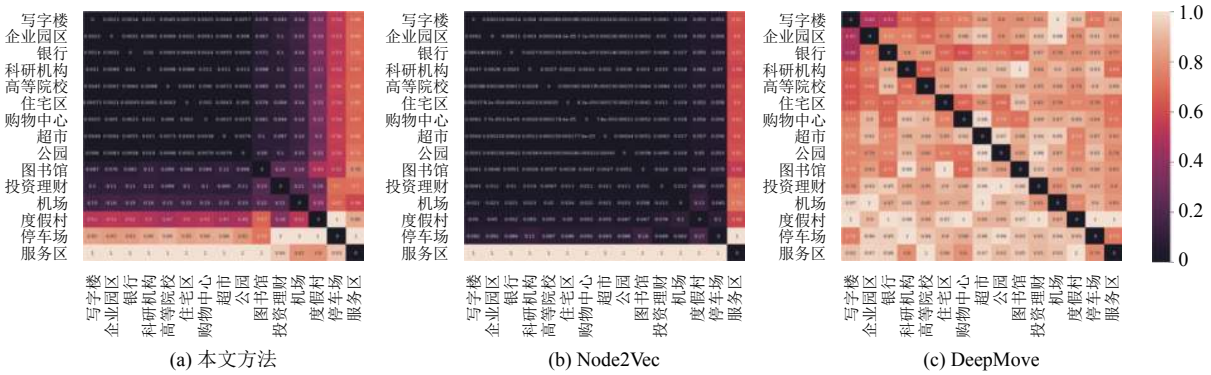
3.2.1 空间位置关联分析

基于使用本文提出的向量化表示方法所生成的 128 维 POI 向量化表示, 计算 POI 之间的相似度 S_{poi}. POI-POI 相似度矩阵如表 2 所示. 具体的, 可发现写字楼与企业园区, 科研机构与高等院校, 住宅区与购物中心的关联度更高, 其向量间的 Cosine 距离也较小. 此外, 由于使用多源交通数据构建网络, 也能使得向量含有居民行为模式特征, 例如机场的 POI 向量与写字楼、高等院校的 POI 向量的距离较小, 而与超市、公园的 POI 向量的距离较大, 表明其关联程度较低. POI 关联性结果可验证本文提出的向量生成方法符合 POI 之间的关联程度愈高, 其间的 Cosine 距离愈小这一规律, 表明使用本文方法能够得到有效且可信度高的向量化表示.

为对比验证本文提出的向量方法的有效性, 将本文方法生成的 POI 向量与其他的已有方法进行对比. 对比方法分别包括 DeepMove[10], Node2Vec[17], 其 POI 的相关性关联热力图如图 5 所示. 对比发现, 使用本文方法得到的向量能够关联 POI 的空间位置特征与居民的行为模式特征; POI 向量间的相似度基本符合先验知识, 并且区分度高, 优于已有方法.

表 2 POI-POI 相关性 (Cosine 距离) 矩阵

POI类型	写字楼	企业园区	银行	科研机构	高等院校	住宅区	购物中心	超市	公园	图书馆	投资理财	机场	度假村	停车场	服务区
写字楼	0.0000	0.0023	0.0014	0.0108	0.0045	0.0007	0.0025	0.0049	0.0057	0.0756	0.0930	0.1393	0.3239	0.5386	0.6786
企业园区	0.0023	0.0000	0.0022	0.0091	0.0069	0.0021	0.0051	0.0093	0.0080	0.0672	0.1027	0.1525	0.3409	0.5300	0.6883
银行	0.0014	0.0022	0.0000	0.0104	0.0065	0.0009	0.0024	0.0055	0.0056	0.0719	0.1008	0.1444	0.3326	0.5324	0.6880
科研机构	0.0107	0.0089	0.0102	0.0000	0.0098	0.0089	0.0110	0.0113	0.0131	0.0983	0.1003	0.1535	0.3135	0.5395	0.6711
高等院校	0.0045	0.0067	0.0064	0.0098	0.0000	0.0043	0.0060	0.0072	0.0093	0.0855	0.0898	0.1456	0.2962	0.5636	0.6760
住宅区	0.0007	0.0021	0.0009	0.0091	0.0043	0.0000	0.0020	0.0043	0.0050	0.0759	0.0944	0.1410	0.3218	0.5415	0.6856
购物中心	0.0025	0.0050	0.0023	0.0110	0.0060	0.0020	0.0000	0.0037	0.0075	0.0814	0.0942	0.1411	0.3230	0.5403	0.6743
超市	0.0049	0.0092	0.0055	0.0114	0.0073	0.0043	0.0038	0.0000	0.0076	0.1015	0.0867	0.1407	0.2959	0.5575	0.6823
公园	0.0060	0.0083	0.0058	0.0139	0.0098	0.0051	0.0079	0.0079	0.0000	0.0904	0.1018	0.1470	0.3150	0.5407	0.7112
图书馆	0.0870	0.0762	0.0815	0.1144	0.0987	0.0865	0.0942	0.1162	0.0992	0.0000	0.2363	0.2366	0.4872	0.5250	0.7807
投资理财	0.1023	0.1114	0.1094	0.1116	0.0992	0.1029	0.1043	0.0949	0.1069	0.2260	0.0000	0.2131	0.2634	0.6961	0.7047
机场	0.1476	0.1593	0.1508	0.1644	0.1548	0.1479	0.1504	0.1483	0.1486	0.2178	0.2051	0.0000	0.3509	0.6700	0.5825
度假村	0.5120	0.5312	0.5185	0.5010	0.4700	0.5035	0.5139	0.4652	0.4752	0.6694	0.3784	0.5238	0.0000	1.0000	0.8628
停车场	0.8514	0.8259	0.8302	0.8624	0.8943	0.8473	0.8596	0.8765	0.8156	0.7213	1.0000	1.0000	1.0000	0.0000	1.0000
服务区	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9437	0.8105	0.8043	0.9322	0.0000



(a) 本文方法 (b) Node2Vec (c) DeepMove

图 5 分别使用本文提出的向量化表示方法、Node2Vec 与 DeepMove 方法生成 POI 向量的相关性热力图

为验证本文提出的多源数据构建大规模交通网络方法与交通网络采样方法的有效性, 设置多组对比实验分别对模型、交通网络数据源、采样方法加以评估. 基于 K-means 聚类方法对 POI 向量进行聚类操作, 其超参数 K 为聚类类目, 可表征为 POI 的类目. 在已有基于兴趣点 (POI) 大数据的研究综述中表明沈阳市有 21 个 POI 类目[18], 由于 POI 类目受多种人为主观分类因素影响, 因此可认为 K 值应在 5~20

范围内浮动. 为避免聚类类目 K 值选取不当对实验结果造成的干扰, 本文将 K 分别设定为 5、10、15、20 以评估不同 POI 聚类类目下的聚类效果, 其中 K 值越大, 类别越细致. 并且每个 K 值分别进行 5 次实验取平均值. 使用平均轮廓系数对 POI 向量聚类效果进行评估. 表 3 给出了使用多种对比模型的实验结果, 其中涉及的对比方法均使用融合了长、短距离轨迹的交通网络. 实验结果表明, 当 POI 聚类类目 K=15、K=20 时, 由于本文提出方法相较于对比方法, 既能够聚合了邻居节点的信息, 又能够利用空间位置自身的特征, 改善 DeepMove[10] 与 Node2Vec[17] 方法仅使用网络中的节点序列特征这一问题, 因此当 POI 分类类目较细致时, 使用本文提出方法生成的向量能够保留更加细粒度的特征, 其聚类效果优于对比方法. 表 4 对比了不同采样方法的实验结果, 实验结果表明在多数 K 值下, 使用本文提出的路网采样方式生成的向量, 其聚类效果都优于使用图神经网络中随机采样方式. 表 5 对比了不同数据源的实验结果, 实验结果表明, 当 K 值较小时, 使用代表长距离出行的公共交通线路所构建的交通网络, 生成的向量聚类效果更好, 但当 K 值逐步增大时, 使用长、短距离融合的交通网络所生成的向量, 其聚类效果比较稳定, 当 K=10、K=15 时, 优于仅使用长距离出行轨迹或短距离出行轨迹构建的交通网络. 因此可推断得出, 当 POI 分类较为粗泛时, 长距离轨迹能区分不同功能区域, 因此长距离轨迹构建的网络效果较好, 当 POI 分类较为细致时, 融合长、短距离的出行轨迹能够捕捉到更加细粒度并且相对完整的出行模式, 因此构建的交通网络所涵盖的信息更加丰富, 效果更优.

表 3 不同向量化方法的实验结果对比 (轮廓系数)

模型	K=5	K=10	K=15	K=20
融合模型	0.5137	0.3254	0.4984	0.3035
Node2Vec	0.3466	0.1579	0.193	0.265
DeepMove	0.7991	0.4988	0.4525	0.2698

表 4 不同采样方法的实验结果对比 (轮廓系数)

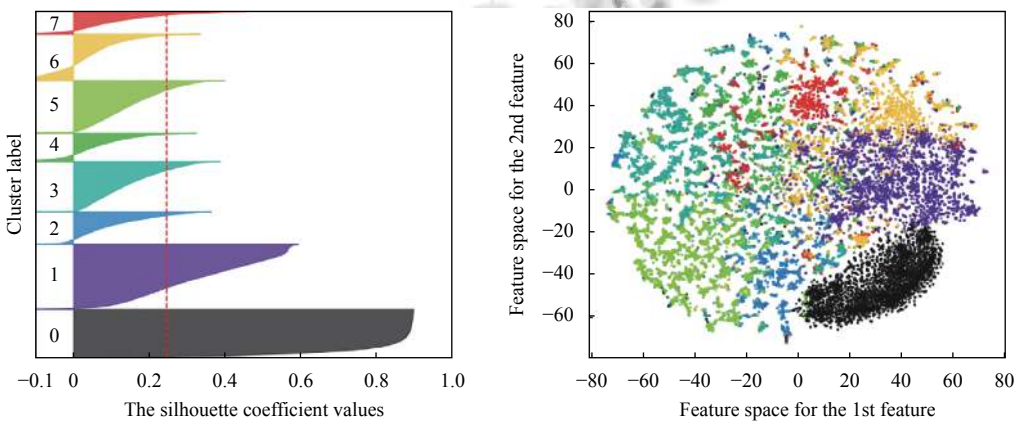
采样方法	K=5	K=10	K=15	K=20
路网采样	0.5137	0.3254	0.4984	0.3035
随机采样	0.4701	0.4369	0.3923	0.2853

表 5 不同数据源的实验结果对比 (轮廓系数)

交通网络数据源	K=5	K=10	K=15	K=20
多源数据-长、短距离融合	0.5137	0.3254	0.4984	0.3035
共享单车轨迹-短距离出行	0.5139	0.4122	0.4013	0.2811
公共交通线路-长距离出行	0.6601	0.6317	0.4488	0.2543

3.2.2 空间位置聚类分析

首先, 利用 K-means (K=8) 对空间位置向量进行聚类分析, 其中每个类别的轮廓系数如图 6(a) 所示. 为了更加直观的反映聚类结果, 使用 t-SNE[19] 方法将空间位置向量从 128 维空间降维至低维空间, 降维后的空间向量化表示的聚类结果如图 6(b) 所示, 不同的颜色表明不同的类. 本文结合低维空间向量化表示的可视化映射图作为一种直观的评价方式. 通过观察图 6 可发现, 使用本文提出的空间向量化表示方法得到的空间位置向量, 映射在低维空间后, 有较为明显的类间边界, 能够验证本文提出的空间向量化表示方法的有效性.



(a) 空间位置向量聚类的轮廓系数分布图 (b) 经过 t-SNE 降维后的空间向量化表示聚类可视化视图

图 6 聚类分析结果

4 结论与展望

本文提出了一种基于图神经网络的空间向量化表示方法. 基于共享单车轨迹数据与公共交通线路数据, 将长、短距离出行轨迹进行匹配连接, 构建大规模交

通网络, 该交通网络能够覆盖多种出行模式. 提出了融合 POI 与轨迹信息的空间向量化表示方法, 综合位置自身的空间特征与其邻域的特征, 并优化节点采样方法, 提高了空间向量化表示的表达能力. 以北京市的共享单车轨迹数据与公共交通路网数据为实例, 经验证本文提出的空间向量化表示方法能够综合空间特征、邻域特征与居民出行模式, 该向量可作为空间特征用于交流流量预测, 交通调度与管理, 地理画像, 位置推荐等实际应用中.

在未来的工作中, 将进一步研究融合多源数据, 例如出租车, 网约车数据, 以构建大规模的交通网络. 以及当网络规模增大时, 如何提升模型性能, 使其能够处理更大规模的交通网络.

参考文献

1 Liu ZD, Li ZJ, Li M, *et al*. Mining road network correlation for traffic estimation via compressive sensing. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(7): 1880–1893. [doi: 10.1109/TITS.2016.2514519]

2 Yan B, Janowicz K, Mai GC, *et al*. From ITDL to Place2Vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, NY, USA. 2017. 1–10.

3 Jin JQ, Xiao ZJ, Qiu Q, *et al*. A GeoHash based Place2Vec model. IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama, Japan. 2019. 3344–3347.

4 Ying R, He RN, Chen KF, *et al*. Graph convolutional neural networks for web-scale recommender systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA. 2018. 974–983.

5 Cai HY, Zheng VW, Chang KCC. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616–1637. [doi: 10.1109/TKDE.2018.2807452]

6 Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 2018, 151: 78–94. [doi: 10.1016/j.knosys.2018.03.022]

7 Chai D, Wang LY, Yang Q. Bike flow prediction with multi-graph convolutional networks. Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, NY, USA. 2018. 397–400.

8 Diao ZL, Wang X, Zhang DF, *et al*. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, HI, USA. 2019. 890–897.

9 Wang YD, Yin HZ, Chen HX, *et al*. Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA. 2019. 1227–1235.

10 Zhou Y, Huang Y. DeepMove: Learning place representations through large scale movement data. Proceedings of 2018 IEEE International Conference on Big Data. Seattle, WA, USA. 2018.

11 Balkić Z, Šoštarić D, Horvat G. GeoHash and UUID identifier for multi-agent systems. Proceedings of the 6th KES international conference on Agent and Multi-Agent Systems. Berlin, Germany. 2012. 290–298.

12 Hamilton W, Ying ZT, Leskovec J. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems. Long Beach, CA, USA. 2017. 1024–1034.

13 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2013. 3111–3119.

14 Ramos J. Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning. Honolulu, HI, USA. 2003. 890–897.

15 Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. 2014. 1532–1543.

16 Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, 20: 53–65. [doi: 10.1016/0377-0427(87)90125-7]

17 Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2016. 855–864.

18 薛冰, 李京忠, 肖骁, 等. 基于兴趣点 (POI) 大数据的人地关系研究综述: 理论、方法与应用. 地理与地理信息科学, 2019, 35(6): 51–60. [doi: 10.3969/j.issn.1672-0504.2019.06.009]

19 Van Der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11): 2579–2605.