

融合时空行为与社交关系的用户轨迹识别模型

张 伟¹⁾ 李 扬¹⁾ 张 吉²⁾ 王建勇³⁾

¹⁾(华东师范大学计算机科学与技术学院 上海 200062)

²⁾(之江实验室 杭州 311000)

³⁾(清华大学计算机科学与技术系 北京 100084)

摘 要 随着位置社交媒体的流行,用户移动行为数据得到极大丰富,推动了用户轨迹识别问题相关研究.该问题着眼于判定目标轨迹所属用户,有助于理解用户移动模式,促进个性化推荐等下游应用.目前已有方法通常尝试采用多分类方法解决该问题.然而,这些研究仍然面临着两个亟待解决的挑战:用户轨迹稀疏和类别数量庞大.其中,前者产生的原因在于用户常常只在社交媒体中发布部分访问地点信息,并且用户兴趣地点分布具有局部性;后者则是因为多分类方法下每个类别即代表一个用户,而用户数量庞大.为应对上述挑战,本文致力于有效利用轨迹时间戳序列和用户社交关系这两类被相关研究忽略的信息.一方面时间戳信息能够缓解轨迹稀疏性,另一方面社交关系可以通过刻画用户(类别)之间相关性帮助用户表示学习.为此,本文提出了融合神经时间过程与图神经网络的新模型 NTPP-GNN (Neural Temporal Point Process with Graph Neural Network),包含空间、时间、社交关系三个模块.空间模块中,双向循环神经网络用来刻画地点间序列性;时间模块中,本文提出双向神经时间过程从正反两个方向捕捉时间连续性,并以此促进轨迹的时间表征;社交关系模块中,图神经网络用于传播和学习用户表示. NTPP-GNN 采用端到端方式学习上述三个模块,以确保模块之间彼此适配.为验证 NTPP-GNN 的有效性,本文在三组数据集 (Foursquare、 Gowalla 和 Brightkite) 上进行实验.结果表明: (1) NTPP-GNN 性能比最好的基准方法在 ACC@1 上平均提高 7.0%; (2) NTPP-GNN 的各个模块对于性能均有贡献; (3) 所提出的双向神经时间过程相比于只考虑先后顺序的传统神经点过程方法能够带来额外提升.

关键词 用户轨迹识别; 循环神经网络; 神经时间过程; 图神经网络; 时空序列; 社交关系
中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2021.02173

A User Trajectory Identification Model with Fusion of Spatio-Temporal Behavior and Social Relation

ZHANG Wei¹⁾ LI Yang¹⁾ ZHANG Ji²⁾ WANG Jian-Yong³⁾

¹⁾(School of Computer Science and Technology, East China Normal University, Shanghai 200062)

²⁾(Zhijiang Lab, Hangzhou 311000)

³⁾(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract With the flourish of location-aware online social platforms, user behavior of mobility has been greatly enriched, which promotes the relevant studies on the user trajectory identification problem. For a given target trajectory, this problem aims to identify a specific user that the trajectory belongs to, which is

收稿日期: 2020-01-03; 在线发布日期: 2020-10-21. 本课题得到国家自然科学基金 (No. 62072182, No. 61702190, No. 61532010)、之江实验室 (No. 2019KB0AB04) 资助. 张 伟 (通信作者), 博士, 副研究员, 计算机学会 (CCF) 会员 (62154M), 主要研究领域为数据挖掘. E-mail: zhangwei.thu2011@gmail.com. 李 扬, 硕士研究生, 主要研究方向为用户个性化建模. E-mail: leeeyang@qq.com. 张 吉, 博士, 高级研究员, 主要研究领域为大数据分析、数据挖掘、信息隐私保护及安全. 王建勇, 博士, 教授, 计算机学会 (CCF) 高级会员, 主要研究领域为数据挖掘.

beneficial for understanding the mobility patterns behind user trajectories and could provide positive influence on a variety of downstream applications, such as personalized recommendation, to name a few. By far, a few existing relevant studies try to utilize multi-class classification methods to tackle the problem. However, these studies still face two main unresolved challenges which need to be addressed in the literature: the sparsity of user trajectories and the large number of categories to be used for classification. Among the two challenges, the reason of the first one is that users tend to choose only a limited number of visited locations to be published in online social medias and their preferred visited POIs are distributed in local regions; the second challenge is caused by the fact that in multi-class classification, each category denotes one user and the number of users is large. To address the above two challenges, this paper aims to effectively utilize the two types of information, i. e., sequences of timestamps in user trajectories and social relations among users, both of which have not been investigated by previous studies for the considered problem. On the one hand, timestamps could be used as additional information to alleviate the sparsity issue. On the other hand, social relations could be leveraged to characterize the correlations between users, which in turn helps user representation learning. To effectively leverage the two types of information for the studied problem, we propose a novel model which couples Neural Temporal Point Process with Graph Neural Network, named NTPP-GNN. This model composes of three modules for the spatial aspect, the temporal aspect, and the social aspect, respectively. In the spatial module, bidirectional recurrent neural networks are exploited to characterize the sequential nature of visited locations contained in trajectories. In the temporal module, we propose a bidirectional neural temporal point process to capture the continuity of time in both forward and backward directions, which could be further employed for facilitating the temporal representation learning of trajectories. In the social relation module, graph neural networks are exploited to encode user social relations into their representations. The whole model NTPP-GNN adopts an end-to-end learning fashion to optimize the above three modules jointly. This could ensure that each learnable module to be compatible with each other for achieving better trajectory identification performance. To validate the effectiveness of NTPP-GNN, we conduct extensive experiments on three widely used available datasets(i. e., Foursquare, Gowalla, and Brightkite). The experimental results indicate that: (1)the proposed NTPP-GNN improves the best-performed baseline w. r. t. ACC@1 by 7.0% in average; (2)each module in NTPP-GNN indeed makes a positive contribution to the final multi-class classification performance; and(3)the developed bidirectional neural temporal point process brings additional performance gains over conventional neural temporal point process approaches that only consider the forward direction in modeling sequential timestamps.

Keywords user trajectory identification; recurrent neural network; neural temporal point process; graph neural network; spatio-temporal sequence; social relation

1 引 言

随着移动互联网时代的深入发展,位置社交网络(Location-aware Social Networks, LSNs)越发受到用户的欢迎.在 LSNs 中,用户可上传个人实时位置信息,并与朋友分享兴趣点(Point-of-Interests, POIs).例如,在基于地点签到的社交网络 Foursquare 中,目前已有 5 千万的日活跃用户参与足迹分享.在中国,以微博为代表的弱社交关系

平台和以微信为代表的强社交关系平台均提供用户上传位置服务.这些社交网络记录了用户大量的实时位置信息,形成了丰富的用户轨迹数据.其中,每条用户轨迹均由用户所访问的地点和时间二元组序列构成.面向用户轨迹的智能分析与挖掘已成为近年来学术界和工业界关注的热点^[1],涉及内容包括地点推荐^[2]、异常轨迹检测^[3]、城市规划^[4]等.这对于提升平台用户体验,优化政府管理效能等方面具有积极意义.

用户轨迹识别正是这一背景下所诞生的新研究问题,强调学习轨迹与用户间相关性,以实现将目标轨迹与已知用户自动关联的目的,因而与其他用户轨迹研究具有本质不同.用户轨迹识别具有重要意义:首先,位置社交平台通过学习识别用户轨迹,可用于推断第三方公司和机构获取的未标记轨迹数据所对应潜在用户.这对于识别罪犯^[5-6]、个性化推荐^[7](注:仅提供偏好信息,可不泄露用户身份)等应用有积极推动作用;其次,对于位置社交平台本身而言,可能存在恶意注册用户或机器人用户.平台需要从用户产生的轨迹中自动识别出这些恶意用户,优化平台环境;此外,用户轨迹识别对于关联同一用户数据也有所帮助,有利于同时捕捉用户短期偏好^[8]和长期偏好^[9].

日常生活中,用户轨迹往往具有一定持续性和重复性^[10].比如,用户总是在办公地点和居住地点间路径上签到.此外,不同用户的轨迹具有一定独特性,能够反映用户的个性化属性^[11].以上特点说明识别轨迹所属用户具有一定可行性.面向轨迹序列的传统建模方法主要包括主题模型^[12-13]、矩阵分解模型^[14]、基于马尔科夫假设的分解机模型^[15,7]等.然而,这些模型对于轨迹的序列性或者主动忽略,或者考虑不足.例如基于马尔科夫假设的分解机模型仅考虑当前访问地点对于后续访问地点的影响.近年来,循环神经网络(Recurrent Neural Networks, RNNs)及其变种^[16]在以自然语言处理为代表的序列建模任务上取得了巨大成功^[17].在轨迹建模方面,实验证明 RNNs 也有着较好的性能^[5-6].因此,目前直接针对用户轨迹识别的代表性研究沿袭了基于 RNNs 的技术路线并将该问题看作多分类任务^[18-19],其中每个分类即代表一个具体用户.

然而,上述方法仍然面临两个主要挑战.第一个挑战为用户轨迹稀疏的问题.真实场景下,用户并不时时刻刻主动上传并分享其位置信息到 LSNs 中,造成访问地点数据存在缺失问题.另一方面,通讯公司搜集的完整位置序列又因为隐私保护等原因而很难直接提供给数据建模方使用.此外,因为用户的个人偏好习惯以及物理世界运动规律^[20],个体访问过的地点在空间区域中分布具有局部性.因此用户轨迹识别方法能够感知的位置序列数据较为稀疏.第二个挑战是类别数量庞大.用户轨迹识别问题可等价于用户多分类问题,其中每个特定类别即代表一个用户,通过多分类方法可判定轨迹所属用户.但是,给定的用户集合往往较大,从而导致类别数过大.这会增加刻画类别表示之间关联性的难度.因而该问题相较于一般多分类问题而言更具挑战性.

为缓解上述挑战带来的局限性,本文致力于从两类被相关研究忽略的信息中提炼知识,达到提升识别效果的目的.第一类为时间序列信息.一般而言,轨迹数据中除了由地点构成的序列外,还包含每个访问地点对应的时间戳信息.但是,前人工作仅利用该信息确定序列中每个地点的相对顺序,并未能充分探索时序信息的价值.相比于地点序列反映的空间特性,时间戳信息构成的序列则体现了轨迹的时间特性.因此,时间序列可作为地点序列的补充信息,对于减轻数据稀疏问题将有所帮助.第二类为社交关系信息.目前的社交平台中,用户之间往往存在连接关系.根据社交关系理论,如果两个用户之间有越多直接相连的边(一阶相似性)^[21]或者更多相同的邻居(二阶相似性)^[22],则两者越相似.反之,两个用户越不同.是以社交关系能够提供用户即类别之间的关联关系,有利于在类别数量较多时产生具有关联性和区分度的类别向量化表示.

为实现上述两类信息与地点序列的有机融合,本文提出了融合神经点过程与图神经网络的新神经模型 NTPP-GNN,其主要包含三个针对空间、时间、社交因素的定制化模块.具体而言,本文利用双向循环神经网络对轨迹中地点序列建模,通过刻画地点序列得到地点相关的整体序列表示.为进一步捕捉时间序列中时间戳自身和相邻时间差所蕴含的连续性,本文基于时间点过程^[23]和循环神经网络提出了双向神经时间点过程.与循环神经时间点过程方法^[24]着眼于序列中未来事件预测不同,双向神经时间点过程强调获得时间序列的整体表示,因此可以从反向时间序列信息建模角度更为充分地利用时间信息.对于用户之间的社交关系, NTPP-GNN 通过图神经网络^[25]对关系建模,利用边的传递性将邻居用户的表示作为目标用户表示更新的依据.通过上述建模学习得到的地点序列表示和时间序列表示将构成轨迹表示,用于计算与用户表示之间的相关性. NTPP-GNN 采用端到端方式学习上述模块并实现目标轨迹的用户推断.

本文的主要贡献点总结如下:

(1) 本文提出利用时间序列和社交关系两种类型信息以应对当前用户轨迹识别方法所面临的用户数据稀疏和类别数量庞大这两方面挑战.

(2) 本文提出了融合神经点过程与图神经网络的新模型 NTPP-GNN,其能够有效建模空间、时间、社交关系三方面信息,以及学习轨迹表示和用户表示之间的关联.

(3) 本文在三组公开的数据集 Foursquare、 Gowalla 和 Brightkite 上测试了模型的性能.实验显

示本文提出的模型性能超越了目前已有的最佳基准方法. 此外, 实验验证了模型融合时间信息或社交关系均对最终性能起到促进作用.

2 相关工作

2.1 用户轨迹建模与识别方法

用户轨迹建模作为研究用户序列行为的关键一环, 已有研究工作相对较多. 早期基于概率主题模型^[12-13]或者矩阵分解模型^[14]的地点推荐研究工作倾向于在独立同分布假设下对单个访问地点行为进行建模. 但上述工作思路忽略了地点访问行为中蕴含的序列性, 未能揭示真实场景下的用户移动行为规律. 为此, 下一访问地点预测和推荐^[26]成为近年来的主流研究方向, 代表性方法包括: (1) 基于马尔科夫假设的分解机模型^[15]. Cheng 等人^[27]以该模型为基础, 额外考虑下一地点到当前地点之间的距离. 具体而言, 将地理空间划分成多个格子区域, 要求候选地点所属格子与用户当前地点所处格子相同或者相邻. 受此启发, Zhang 等人^[7]进一步探索了时间因素、损失函数选择如何影响预测性能. 然而, 这些模型均假定下一地点只受当前地点影响, 未能充分利用轨迹序列信息. (2) 基于序列建模的神经网络方法. 该类型方法能够捕捉轨迹的序列性, 效果上优于前一类型方法, 因此逐渐成为轨迹建模的代表性方法. 文献^[5-6,28]均以循环神经网络为基石从轨迹中学习用户偏好.

与上述工作中推断对象为单一地点不同, 用户轨迹识别中推断对象为轨迹归属用户, 强调学习轨迹整体表示. 因而该研究问题与其他用户轨迹建模问题有显著不同, 属于近年来新提出的问题, 其研究工作相对较少. Gao 等人^[18]对该问题进行了形式化阐述, 并提出使用基于双向和堆叠循环神经网络的多分类模型. 其中, 单个地点的表示通过类似词嵌入表示学习方式 Word2Vec^[29]获得. 文献^[19]则进一步利用变分自编码器 (Variational Auto-Encoder, VAE)^[30]的数据重构过程, 将未标记用户的轨迹数据纳入其半监督训练模式中, 此举对于克服数据稀疏性有一定帮助.

与前人用户轨迹识别研究不同, 本文从引入时间序列和社交关系角度出发, 强调应对用户数据稀疏和类别数量庞大这两大挑战.

2.2 时间点过程及其在序列建模中的应用

时间点过程是一种对事件发生可能性以及发生时间进行刻画概率统计方法^[23], 其可以运用于

涉及时序建模的多种应用中. 比如, 在地理学领域, 事件可以表示为地震的发生. 时间点过程可以帮助地震监测机构预测未来某时某地发生地震的可能性^[31]. 在社交媒体领域, 事件可由一系列用户在不同时间产生的行为组成, 这些行为包含了用户对他人生成内容的转发和评价等^[32].

近年来提出的神经时间点过程将神经网络强大的函数拟合能力与时间点过程结合, 用于描述其中关键的条件强度函数^[33]. 神经时间点过程在基于用户轨迹的下一地点预测任务上取得了一定成功^[24,34-35], 其最大优势在于能够捕捉时间戳及相邻时间差的取值连续性, 并能够同时预测行为产生的具体时间. 不失一般性, 本文以文献^[24]所提出的神经时间点过程 RMTTP (Recurrent Marked Temporal Point Processes) 为例介绍其核心计算过程, 并在后续章节中对该方法进行扩展以适配用户轨迹识别问题.

具体来讲, RMTTP 利用循环神经网络对事件时间序列建模. 假设序列中第 j 个事件发生时间为 t_j , 此时循环神经网络隐藏层的输出表示为 h_j , 其形式化描述如下:

$$h_j = W^x x_j + W^t t_j + W^h h_{j-1} + b^h \quad (1)$$

其中, t_j 为时间 t_j 对应的特征向量表示, x_j 为事件 j 的特征向量表示, h_{j-1} 为上一时刻隐藏层的输出, W^x 、 W^t 、 W^h 和 b^h 为待优化的模型参数矩阵和向量. 此时, RMTTP 对应的条件强度函数 $\lambda^*(t)$ 表示为:

$$\lambda^*(t) = \exp(vh_j + w(t - t_j) + b) \quad (2)$$

其中, v 为参数向量, w 和 b 均为参数标量. 指数算子 $\exp$ 用于确保强度函数取值为正, 以满足时间点过程对于强度取值的约束. 指数项中第一项表示截至事件 j 发生为止所有过去事件对未来事件发生的影响. 第二项刻画事件 j 发生时间与未来事件发生时间的差值对于其发生的影响. 最后一项代表条件强度函数的基值, 衡量所有事件发生可能性的基准大小.

借助条件强度函数的表达式, 可以定义下一个事件发生时间的概率密度函数表示如下:

$$f^*(t) = \lambda^*(t) \exp\left(-\int_{t_j}^t \lambda^*(\tau) d\tau\right) \quad (3)$$

通过在已知事件集合上最大化上述密度函数, 可以求得神经时间点过程模型的参数.

本文在神经时间点过程基础之上, 提出融合双向时间序列信息, 并与地点序列和社交关系建模部

分结合, 以提升用户轨迹识别效果。

2.3 图神经网络及其应用

近年来兴起的图神经网络是针对图类型数据建模的有效方法, 通过节点表示传播与学习可将图中节点之间关联性嵌入到节点低维表示中。目前使用最广泛的图神经网络包含图卷积神经网络 (Graph Convolutional Neural Network, GCN)^[36]和图注意力神经网络 (Graph Attention Network, GAT)^[25]等。图神经网络在实际问题中有广泛的应用, 包括社交关系网络、分子结构图等。Wu 等人^[37]为此做了较为详细的综述, 可供参考。

特别地, 面向异构网络的表示学习^[38]从邻近关系保持法、消息传递法、关系学习法三方面建模节点和边的多样性, 与本文所研究的用户、时间、空间等异构信息有一定联系。在面向时空数据的异构图表示学习方面, 最近一系列研究工作^[39-41]主要利用图神经网络学习路网中关键节点间的空间邻近关系, 用于实时交通流量预测。然而上述方法旨在捕捉用户群体反馈, 不适用于诸如本文所研究的用户个体建模任务。在用户个体建模方面, 研究^[42-43]分别从异构节点和异构关系角度提出定制化的图神经网络方法建模用户个体历史行为, 以服务于个性化推荐应用; Jin 等人^[44]则在所提出异构图表示学习方法中融入用户个体语义表征, 服务于用户社区发现。与之不同的是, 用户轨迹识别问题可看作对于给定时空序列进行多分类, 因而这些方法不能直接用于该问题。

本文采用 GAT 学习用户社交关系, 并与其他建模部分融合。用户社交关系是一种典型的图结构数据, 图中的节点表示用户, 边通过取值 1 或者 0 表示相应用户之间是否存在关系, 例如朋友关系等。相对于 GCN 而言, GAT 能够在已知用户之间具有关系的基础之上自动学习该关系的强度, 是对社交网络中用户边的原始权重的一种细化。GAT 由多个图注意力层 (Graph Attention Layer, GAL) 构成, 其中图注意力层的输入是一系列节点的向量表示, 输出是传播和更新之后节点的向量表示。图注意力神经网络一般可以叠加多个图注意力层来实现更强的表征能力。

值得注意的是, 本文并未选择异构图神经网络方法整体建模时空、用户、社交关系等异构信息。这主要基于以下两方面考虑: 首先, 利用点过程建模时间信息是本文的一个研究重点, 然而如何在图神经网络中引入时间点过程方法还需要认真探索; 其次, 由于本文需要从时间地点轨迹序列层面进行用

户轨迹识别, 因此是否需要在异构图中引入序列层面对应节点, 亦或是在异构图表示学习基础之上通过序列神经网络学习轨迹表示, 仍然需要深入研究。基于上述考虑, 本文将设计面向用户轨迹识别问题的异构图神经网络作为未来工作之一。

3 问题定义与数据分析

本节首先对全文使用符号类型进行约定, 并形式化定义用户轨迹识别问题, 包括必要的符号含义说明。之后对轨迹时间信息和用户社交关系进行简要分析, 以阐明本文考虑时间序列和社交关系信息动机的合理性。

3.1 问题定义

给定用户集合 $U = \{u_1, u_2, \dots, u_{|U|}\}$ 和地点集合 $P = \{p_1, p_2, \dots, p_{|P|}\}$, 其中 $|U|$ 和 $|P|$ 分别表示用户集合和地点集合的大小。对于特定用户 u_i , 可假定其关联的已知轨迹数表示为 c_i , 其中第 z 条轨迹表示为 $S_z^i = \{R_1, R_2, \dots, R_{|S_z^i|}\}$, $|S_z^i|$ 对应该轨迹的长度。 R_m 指代该轨迹序列中的第 m 个记录, 由二元组构成, 可表示为 $R_m = \{p_m, t_m\}$, 其中 t_m 为时间戳对应数值。另外, 用户之间的社交关系图可定义为 $G = \{U, E\}$, E 表示用户间形成边的集合。

给定所属用户未知的目标轨迹 S , 用户轨迹识别问题希望学习得到一个模型 φ , 能够自动从用户集合 U 中推断该轨迹所属的具体用户 u , 即 $\varphi: S \rightarrow u$ 。需要注意的是, 本文规定 S 的所属用户来自集 U 。对于用户不属于 U 的情况, 可通过添加一个未知用户作为新类别对集合 U 进行扩展。相关研究和实验留作未来工作。

3.2 数据分析

本节首先在 Foursquare^[45]和 Gowalla^[46]两组公开数据上进行分析。这两组数据也将作为实验数据使用, 详细数据统计信息在实验设置章节中介绍。

首先分析用户轨迹蕴含的时间序列信息特点。图 1 描述了 Foursquare 中两个用户的地点访问频度与每天不同时间段的关系。其中, 横轴表示用户每天的签到时间; 纵轴表示该时间段签到频率, 当两名用户签到频率相同时图中标记重叠。可以看出, 用户 A 在 9:00-12:00、14:00-18:00 之间发布的地方访问记录频度低, 而在 12:00-14:00 以及 18:00-24:00 之间则相对较大。与之对应的是, 用户 B 发布的地点访问频度在不同时间段较为分散。因此可以直观分析, 用户 A 可能是一位白领, 而用户 B 则可能是一位自由职业工作者。

进一步从 Foursquare 数据集中随机抽取 500 个用户呈现其访问地点之间平均时间间隔分布情况. 如图 2 所示, 横坐标表示平均时间间隔; 纵坐标表示在该时间间隔下的用户数量. 可以发现, 不同用户访问地点的时间间隔差异也较为明显.

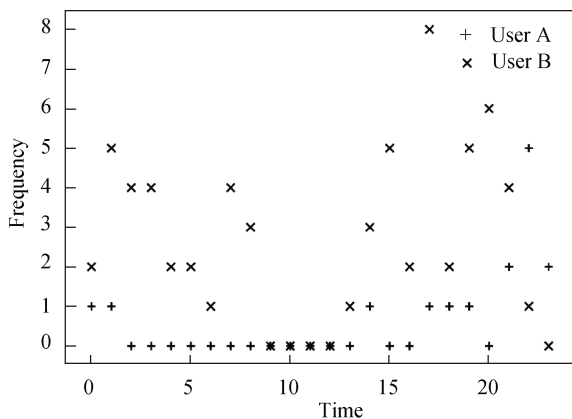


图 1 用户不同时间段访问地点频度情况 (其中, 纵轴对应频度, 无单位; 横轴对应单位为小时)

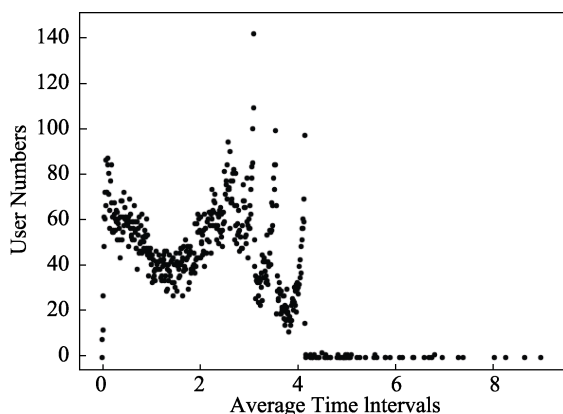


图 2 用户相邻两次访问地点平均时间间隔分布 (其中, 纵轴对应频度, 无单位; 横轴对应单位为小时)

其次分析用户之间拥有社交关系与否对于其访问地点行为的影响. 由于本文所采用的真实数据中用户之间的关系强度为二元取值 (即是否具有关系), 因此本节遵照数据自身特点, 简要比较朋友间访问地点集合的相似度和非朋友间访问地点集合的相似度. 地点集合相似度的计算方式如下: 根据任意给定两个用户 u_a 和 u_b 的已知轨迹, 可提取出他们访问地点的集合分别为 P_a 和 P_b . 利用如下公式计算两者之间 Jaccard 相似度:

$$Jaccard = \frac{|P_a \cap P_b|}{|P_a \cup P_b|} \quad (4)$$

按所有朋友组成的用户对和所有非朋友组成的用户

对分别计算平均 Jaccard 相似度, 可得表 1 所示结果. 据此可以得出, 具有朋友关系的用户之间所访问地点集合相似度更高.

表 1 用户访问地点集合 Jaccard 相似度

数据集	Foursquare	Gowalla
朋友间相似度	0.019	0.008
非朋友间相似度	0.010	0.001

通过上述数据分析, 本节验证了轨迹数据蕴含的两个特点: (1) 不同用户访问地点行为所呈现的时间序列特征不尽相同; (2) 具有社交关系的用户所访问的地点对象相似度相对更高. 上述两方面真实数据分析结果对于本文选择建模用户时间序列和用户社交关系具有指导意义. 特别地, 本文提出了双向循环神经网络过程模型, 能够刻画时间维度取值的连续性, 以充分利用轨迹中的时间序列信息; 采用图神经网络建模用户之间关系, 以此捕捉用户之间的关联性, 增强每个用户代表的类别表征能力.

4 用户轨迹识别模型 NTPP-GNN

本节将给出模型 NTPP-GNN 的详细数学描述. 图 3 展示了 NTPP-GNN 的模型结构图, 其核心主要由三部分组成: 访问地点序列建模部分、访问时间序列建模部分和用户社交关系建模部分. 其中, 前面两部分通过空间和时间维度得到轨迹整体表示, 第三部分得到用户表示. 最后, NTPP-GNN 基于提出的混合优化目标实现三个部分的端到端学习. 接下来本节以用户 u_i 的第 z 条轨迹 S_z^i 为例, 介绍 NTPP-GNN 的执行过程.

4.1 访问地点序列建模

已知轨迹 S_z^i 对应的访问地点序列为 $\{p_1, p_2, \dots, p_{|S_z^i|}\}$. 对于其中每个地点, 采用 one-hot 编码的方式作向量量化表示. 例如对于第 m 个地点 p_m ($m \in \{1, 2, \dots, |S_z^i|\}$), 其向量化表示为 $\mathbf{p}_m \in \{0, 1\}^{|P|}$. 该向量仅有一个维度取值为 1, 对应地点在整个地点词典中的位置. 令 $\mathbf{E}^P \in \mathbb{R}^{|P| \times K}$ 为所有地点向量表示构成的可学习表示矩阵, K 为向量维度大小 (K 具体取值见表 3). 基于查表操作可得到地点 p_m 的低维实数向量表示:

$$\mathbf{e}_m^P = \mathbf{p}_m \mathbf{E}^P \quad (5)$$

通过学习得到的上述向量能够表征地点之间的相关性, 突破 one-hot 编码表示衡量相关性等瓶颈.

本文采用的双向循环神经网络由前向和后向两

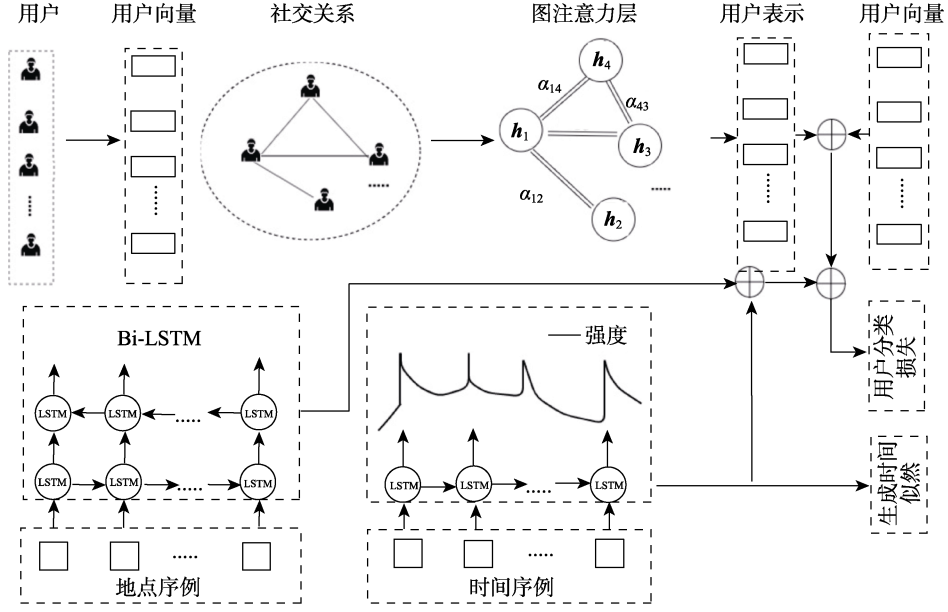


图 3 模型 NTPP-GNN 结构示意图

个 LSTM 网络组成, 表示如下:

$$\overrightarrow{h}_m^P = \text{LSTM}(\overrightarrow{h}_{m-1}^P, e_m^P) \quad (6)$$

$$\overleftarrow{h}_m^P = \text{LSTM}(\overleftarrow{h}_{m+1}^P, e_m^P) \quad (7)$$

根据上述公式, 可以得到代表两个方向的序列整体表示 $\overrightarrow{h}_{|S_i|}^P$ 和 $\overleftarrow{h}_1^P$. 通过简单加和操作对其融合:

$$h_{i,z}^P = \overrightarrow{h}_{|S_i|}^P + \overleftarrow{h}_1^P \quad (8)$$

因此, $h_{i,z}^P$ 融合了两个方向的表示, 能够更好地表征地点序列, 对于针对序列级别的任务有所帮助.

4.2 访问时间序列建模

对于轨迹 S_z^i 而言, 其对应的访问时间序列 $\{t_1, t_2, \dots, t_{|S_i|}\}$ 由时间戳信息构成, 相邻两次访问的时间差可直接计算得到. 如前所述, 本文采用双向神经时间过程对时间序列建模, 意在取值连续的时间信息做精细化建模. 具体而言, 为在输入端引入时间信息, 本文首先对时间取值进行离散化, 即将每个时间戳取值映射到每天以小时为单位的不同时间区间中. 例如用户发布访问某个地点的时间戳为 2019 年 12 月 28 日上午 10 点 30 分, 则该时间戳映射到 10 点至 11 点这个区间上. 此时, 时间区间个数为 24. 为每个区间关联一个可学习的向量化表示, 则可形成一个时间表示矩阵 $E^T \in \mathbb{R}^{24 \times K}$ (K 具体取值见表 3), 并通过如下公式得到时间戳 t_m 的特征化表示:

$$e_m^T = \text{map}(t_m)E^T \quad (9)$$

其中, $\text{map}(t_m)$ 通过映射得到 one-hot 编码表示.

采用如公式(6)和公式(7)所示方式, 将序列中不同时间表示从正、反两个方向分别关联, 可得:

$$\overrightarrow{h}_m^T = \text{LSTM}(\overrightarrow{h}_{m-1}^T, e_m^T) \quad (10)$$

$$\overleftarrow{h}_m^T = \text{LSTM}(\overleftarrow{h}_{m+1}^T, e_m^T) \quad (11)$$

相比于传统循环神经时间过程方法只考虑正向时间序列建模, 本文从正、反两方向同时进行时间序列建模, 其动机在于本文并不强调利用过去行为事件未来行为时间进行预测, 而是侧重于更加全面地利用时间信息学习时间序列表征上. 为此, 本文定义如下两个时间点过程条件强度函数:

$$\overrightarrow{\lambda}^*(t) = \exp(\overrightarrow{v}^T \overrightarrow{h}_m^T + \overrightarrow{w}(t - t_m) + \overrightarrow{b}) \quad (12)$$

$$\overleftarrow{\lambda}^*(t) = \exp(\overleftarrow{v}^T \overleftarrow{h}_m^T + \overleftarrow{w}(t_m - t) + \overleftarrow{b}) \quad (13)$$

其中, 上述公式的 $\exp$ 函数中第一部分代表两个向量表示做内积运算, 最后得到条件强度标量大小. 公式(13)指代反向时间点过程, 其时间戳取值满足 $t < t_m$. 此时, 关于时间取值的连续性得以被考虑.

上述两式代入公式(3)中所示的概率密度函数, 通过积分可得如下计算式 (注意: 反向点过程积分时积分号中 t 和 t_m 的位置需要调换顺序):

$$\overrightarrow{f}^*(t) = \exp(\overrightarrow{v}^T \overrightarrow{h}_m^T + \overrightarrow{w}(t - t_m) + \frac{1}{w} \exp(\overrightarrow{v}^T \overrightarrow{h}_m^T + \overrightarrow{w}(t - t_m) + \overrightarrow{b})) - \frac{1}{w} \exp(\overrightarrow{v}^T \overrightarrow{h}_m^T + \overrightarrow{w}(t - t_m) + \overrightarrow{b}) \quad (14)$$

$$\overleftarrow{f}^*(t) = \exp(\overleftarrow{v}^T \overleftarrow{h}_m^T + \overleftarrow{w}(t_m - t) + \frac{1}{w} \exp(\overleftarrow{v}^T \overleftarrow{h}_m^T + \overleftarrow{w}(t_m - t) + \overleftarrow{b})) - \frac{1}{w} \exp(\overleftarrow{v}^T \overleftarrow{h}_m^T + \overleftarrow{w}(t_m - t) + \overleftarrow{b}) \quad (15)$$

从形式上看,公式(14)和公式(15)十分类似,但由于前向和后向 $\mathbf{h}_m^T$ 所表征的时间子序列截然不同,因而两式具有本质不同.由于针对时间点过程的优化涉及到最大化上述公式所代表的已知事件在特定时间发生的可能性,因此前向和后向 $\mathbf{h}_m^T$ 的学习过程受到连续时间戳信息的指导,得到的表示能够更有效地捕捉时间信息.

采用类似公式(8)的方式,时间序列整体表示可由如下方式获得:

$$\mathbf{h}_{i,z}^T = \overrightarrow{\mathbf{h}_{i,z}^T} + \overleftarrow{\mathbf{h}_i^T} \quad (16)$$

4.3 用户社交网络关系建模

本文采用图注意力网络 GAT 对用户社交关系图进行建模.首先依然采用查表方式从全体用户表示矩阵 $\mathbf{E}^U \in \mathbb{R}^{U \times K}$ 中获得特定用户 u_i 的初始向量化表示 $\mathbf{e}_i^U$.同时,根据社交关系网络中用户之间的连接关系,确定每个用户的邻居用户集合,例如 N_i 对应用户 u_i 的邻居集合(注:用户本身也包含在其邻居中).

GAT 通过注意力机制确定邻居用户之间边的权重大小.具体而言,对于目标用户 u_i 和其邻居,对应权重系数 α_{ij} 计算公式如下:

$$\alpha_{ij} = \frac{\exp(\text{Leaky ReLU}(\mathbf{w}^G[\mathbf{W}^G \mathbf{e}_i^U \parallel \mathbf{W}^G \mathbf{e}_j^U]))}{\sum_{j \in N_i} \exp(\text{Leaky ReLU}(\mathbf{w}^G[\mathbf{W}^G \mathbf{e}_i^U \parallel \mathbf{W}^G \mathbf{e}_j^U]))} \quad (17)$$

其中, $[\cdot \parallel \cdot]$ 表示任意两个向量的拼接,LeakyReLU中 α 设为 0.5.基于上式计算得到的用户 u_i 与所有邻居用户的边权重,采用如下公式更新 u_i 表示:

$$\tilde{\mathbf{e}}_i^U = \sigma \left(\sum_{j \in N_i} \alpha_{ij} \mathbf{e}_j^U \right) \quad (18)$$

其中, σ 为 ELU 激活函数.为从多方面刻画用户之间的相关性表示,本文沿袭 GAT 采用的多头注意力机制,此时权重系数 α_{ij} 扩展为 $\alpha_{ij}^n (n \in \{1, 2, \dots, Q\})$,其中 Q 代表头部个数.用户节点表示更新公式重新定义为

$$\tilde{\mathbf{e}}_i^U = \sigma \left(\frac{1}{Q} \sum_q \sum_{j \in N_i} \alpha_{ij}^q \mathbf{e}_j^U \right) \quad (19)$$

上述更新过程对应图神经网络中的一次传播,可表示为 $\tilde{\mathbf{e}}_i^U = \text{GAT}(\mathbf{e}_i^U)$.多次传播可在一次传播的基础上自然的扩展,即 $\mathbf{e}_{i(x)}^U = \text{GAT}(\mathbf{e}_{i(x-1)}^U)$,其中 x 对应第 x 次传播.给定传播总次数为 X ,则用户基于社交关系的传播表示为 $\mathbf{e}_{i(X)}^U$.通过将用户社交表示与用户初始化向量化表示结合,本文定义用户最终表示为如下形式:

$$\mathbf{w}_i^U = \mathbf{e}_i^U + \mathbf{e}_{i(X)}^U \quad (20)$$

4.4 模型优化目标

上述三部分建模可以得到两种不同角度的轨迹表示,包括地点序列表示 $\mathbf{h}_{i,z}^P$ 、时间序列表示 $\mathbf{h}_{i,z}^T$,以及用户表示 $\mathbf{w}_i^U$.基于这些表示,本文定义该轨迹属于用户 u_i 的概率为

$$p(u_i | \mathbf{S}_z^i) = \text{Soft max}((\mathbf{h}_{i,z}^P + \mathbf{h}_{i,z}^T) \cdot \mathbf{w}_i^U) \quad (21)$$

通过上述公式可以看到,通过双向神经时间点过程建模得到的时间序列表示 $\mathbf{h}_{i,z}^T$ 作为地点序列表示 $\mathbf{h}_{i,z}^P$ 的补充.对于新给定轨迹,通过上述公式可计算其属于每个用户的概率,并选择概率最大的类别所代表的用户作为识别结果.

为实现面向 NTPP-GNN 的端到端训练方式,本文设计了混合优化目标函数,包含两部分内容:(1) 轨迹所属用户识别损失函数;(2) 基于时间生成概率似然的损失函数.针对第一部分,本文采用交叉熵函数衡量识别类别概率和真实类别之间的差异:

$$\text{loss}_c = \sum_{u_i \in U} \sum_{z=1}^{c_i} y_{i,z} \log p_{i,z} \quad (22)$$

其中, $y_{i,z}$ 为真实分类结果所对应的二值化向量.向量中真实用户对应类别取值为 1,其余皆取值为 0.因而可看作是真实分类的概率分布. $p_{i,z}$ 代表 NTPP-GNN 估计用户概率分布.

对于时间部分,基于公式(14)和公式(15)中给出概率密度函数,定义如下关于时间生成概率似然的损失函数:

$$\text{loss}_t = \sum_{u_i \in U} \sum_{z=1}^{c_i} \sum_{m=1}^{|\mathbf{S}_z^i|} -\log \overrightarrow{f^*}(t_m) - \log \overleftarrow{f^*}(t_m) \quad (23)$$

基于此,本文最终的混合优化目标定义如下:

$$\text{loss} = \text{loss}_c + \lambda \cdot \text{loss}_t \quad (24)$$

其中, λ 为调节两部分损失函数相对权重的超参数.

算法 1 总结了学习 NTPP-GNN 模型参数的主要步骤.需要注意的是,由于该学习算法属于小批量随机梯度下降算法,因此支持增量式更新.具体而言,当新轨迹数据出现时,可以组成的小批量轨迹集合,用于增量式优化 NTPP-GNN 模型.这对于应对历史数据的过期问题有一定帮助.

算法 1. NTPP-GNN 学习算法.

输入: 训练轨迹集合 S^u , 用户集合 U , 地点集合 P , 用户社交关系图 G , 预设迭代次数 $iter\text{-}num$.

输出: 训练好的 NTPP-GNN 模型参数

1. 随机初始化 NTPP-GNN 模型参数
2. FOR $I=1$ TO $ITER\text{-}NUM$ DO

3. 从用户集合 U 中抽样得到一个用户子集
4. 从训练集合 S^u 中提取上述用户子集对应的轨迹集合 B ;
5. 针对轨迹集合 B , 由公式(24)计算优化目标函数值;
6. 基于梯度下降法(例如 Adam)优化上述目标, 并更新 NTPP-GNN 模型参数;
7. END

4.5 方法复杂度分析

NTPP-GNN 的训练时间消耗主要由地点序列建模、时间序列建模、社交网络关系建模三部分构成, 其中前两部分时间占比较大. 由于前人相关方法未考虑时间和社交关系部分, 因此只有地点序列部分需要耗费时间. 以 Bi-TULER 为例, NTPP-GNN 的训练时间消耗约为其 2 倍多一点.

NTPP-GNN 的测试时间消耗与训练时间消耗的主要不同之处为不再需要运行 GAT 部分学习用户社交关系, 而是直接使用训练得到的用户最终表示即可. 具体而言, 通过单块 GPU 显卡(版本: GeForce GTX 1080 Ti)计算 NTPP-GNN 在本文所使用测试数据集上的时间消耗, 可以发现平均每条轨迹识别所需时间在 3ms 以内, 约为 Bi-TULER 的 2 倍左右, 在可接受范围之内.

5 实验设置

本节主要讨论实验中所用数据集、基准对比方法、评价指标以及模型实现这几方面内容.

5.1 实验数据集

本文基于 Foursquare^[45]、Gowalla^[46]和 Brightkite^[46]三个公开数据集进行实验. Foursquare 原始数据集包含从 2012 年 4 月到 2014 年 1 月之间的用户签到数据; Gowalla 中用户签到行为时间范围为 2009 年 2 月至 2010 年 10 月; Brightkite 原始数据集则涉及 2008 年 4 月到 2010 年 10 月的用户签到行为. 三个数据集中均包含用户之间的社交关系. 实验数据集为从原始数据集中导出的子集, 要求每名用户的每日访问序列长度不小于 3, 朋友数量不少于 3, 涉及的天数不少于 5 天. 实验数据的基本统计信息如表 2 所示.

表 2 实验数据基本统计信息

条目	Foursquare	Gowalla	Brightkite
用户数量	807	1052	1500
总样本数量	21336	18701	25394
平均访问天数	26	17	17
平均朋友数量	4	6	6

为验证不同方法效果, 本文将数据集中 60% 作为训练集, 20% 作为验证集, 剩下 20% 作为测试集.

5.2 对比方法

实验主要从两个最相关工作^[18,19]中选取基准方法用于对比:

- TULER-GRU, TULER-LSTM, TULER-LSTM-S^[18]: 三个模型均由文献^[18]提出. 其中, TULER-GRU 表示利用 GRU 捕捉地点序列信息, TULER-LSTM 利用 LSTM 建模地点序列, TULER-LSTM-S 则表示使用堆叠 LSTM;
- Bi-TULER^[18]: 该模型采用双向 LSTM 学习地点序列表示, 以此作为多用户分类部分输入;
- TULVAE^[19]: 该模型利用 VAE 建模用户访问地点序列. 编码器与解码器均使用 LSTM 建模, 编码器的输入与基于循环神经网络的输入相同, 输出为高斯分布的均值和方差. 解码器的输入由高斯分布采样得到. 通过对该向量进行解码, 可以识别用户的身份. TULVAE 模型还能利用未标记数据进行半监督训练.

5.3 评价指标

用户轨迹识别方法致力于判断每条轨迹最可能归属的用户. 因此本文沿用文献^[18-19]采用的评价指标, 包括 ACC@1、ACC@5 和 Macro-F1. 其中, ACC@K 定义为

$$ACC@K = \frac{\text{产生的}K\text{个候选中包含正确用户分类的轨迹数}}{\text{总的轨迹数}}$$

选择较小 K 值的原因在于, 每一条轨迹只对应一个正确分类结果, 需要关注靠前候选.

Macro-F1 用于刻画各个类别平均的 $F1$ 分值, 代表准确率和召回率融合后的结果, 其定义为

$$F1 = \frac{2 * P@1 * R@1}{P@1 + R@1},$$

$P@1$ 表示各个类别关于第一个分类结果的精确率, $R@1$ 则对应召回率.

此外, 本文还考虑 Macro- P (宏精确率) 和 Macro- R (宏召回率) 两个指标, 其可通过对所有位置对应的精确率 (P) 和召回率 (R) 加和平均获得

$$Macro-P = Mean \left(\sum_k P@k \right),$$

$$Macro-R = Mean \left(\sum_k R@k \right).$$

5.4 模型实现

各个对比方法以及所提出模型 NTPP-GNN 均涉及一定量超参数. 本文根据不同超参数下相应方法在验证集上的性能效果差异, 选择合适的超参数用于验证模型在测试集上效果. 本文使用 Adam^[47] 作为优化器, 并使用早停 (early stopping) 技术避免模型过拟合. 经过参数选择, NTPP-GNN 的关键超参数设置如表 3 所示.

表 3 超参数设置

条目	数值
小批量大小	8
用户向量维度	300
地点向量维度	300
时间向量维度	300
学习率	1e-4
注意力头数 (N)	3
GAT 层数 (X)	2
相对权重 ²	0.5

6 实验结果

6.1 模型结果对比

本节对 NTPP-GNN 和基准方法进行比较, 表 4、表 5 和表 6 分别对应 Foursquare 测试集、Gowalla 测试集和 Brightkite 测试集上结果.

首先, 通过对比 TULER-LSTM 和 TULER-GRU 可以发现, 基于 LSTM 的地点序列建模效果在三个数据集上都显著优于基于 GRU 的方式, 说明该任务上模型复杂度更高的 LSTM 可以发挥其特点.

其次, 实验结果显示 TULVAE 性能相比其他基

表 4 测试集 Foursquare 中不同方法对比

方法	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1
TULER-GRU	0.332	0.468	0.331	0.326	0.296
TULER-LSTM	0.366	0.474	0.346	0.341	0.333
TULER-LSTM-S	0.368	0.478	0.344	0.338	0.336
Bi-TULER	0.381	0.513	0.352	0.345	0.349
TULVAE	0.358	0.476	0.341	0.332	0.322
NTPP-GNN	0.409	0.539	0.376	0.368	0.366

表 5 测试集 Gowalla 中不同方法对比

方法	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1
TULER-GRU	0.208	0.321	0.183	0.161	0.151
TULER-LSTM	0.234	0.363	0.202	0.187	0.178
TULER-LSTM-S	0.238	0.366	0.205	0.186	0.175
Bi-TULER	0.251	0.375	0.213	0.199	0.202
TULVAE	0.235	0.357	0.194	0.186	0.180
NTPP-GNN	0.273	0.406	0.232	0.224	0.215

表 6 测试集 Brightkite 中不同方法对比

方法	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1
TULER-GRU	0.530	0.633	0.464	0.466	0.437
TULER-LSTM	0.543	0.644	0.481	0.479	0.441
TULER-LSTM-S	0.546	0.651	0.483	0.477	0.443
Bi-TULER	0.556	0.663	0.494	0.489	0.466
TULVAE	0.536	0.641	0.473	0.475	0.434
NTPP-GNN	0.583	0.681	0.515	0.509	0.484

准方法并没有体现出优势, 仅优于 TULER-GRU. 说明 VAE 的建模的方式虽然能对不确定性做估计, 但针对此任务的序列表示学习并没有明显优势. 进一步对比所有基准方法可以看出, Bi-TULER 方法的效果最好, 超过了堆叠 LSTM 的方法. 上述对比验证了双向建模对于轨迹序列整体表征的重要性.

最后, 通过本文模型 NTPP-GNN 与其他方法总体比较, 体现了 NTPP-GNN 的优势. 具体而言, 三个测试集上的平均效果显示, NTPP-GNN 在 ACC@1、Macro-P、Macro-R 上相比于最好基准方法效果分别相对提升了 7.03%、6.67%、6.69%, 在 Macro-F1 上也有 4.7% 的相对提升.

6.2 模型结构有效性分析

NTPP-GNN 的特色在于融入了面向时间建模的神经点过程方法和面向社交关系建模的图神经网络方法. 为验证上述方法所起作用, 本小节针对该模型进行擦除实验. 其中, NTPP-GNN w/o G 指代从模型中去除社交关系建模部分, NTPP-GNN w/o T 表示从模型中去除时间建模部分.

表 7、表 8 以及表 9 分别展示了擦除实验在 Foursquare、Gowalla 和 Brightkite 测试集上的结果. 首先分别将 NTPP-GNN w/o G 和 NTPP-GNN w/o T 与最好的基准方法 Bi-TULER 进行比较, 通过对比结果可以得出考虑时间序列信息和社交关系信息中的任一方面均能提升结果的结论. 其次, 分别将 NTPP-GNN w/o G 和 NTPP-GNN w/o T 与完整模型作比较, 可以发现完整模型相比其任一子模型均有提高, 验证了时间序列建模和社交关系建模融合的有效性.

表 7 测试集 Foursquare 中擦除实验结果对比

方法	ACC@1	ACC@5	Macro-F1
Bi-TULER	0.384	0.512	0.349
NTPP-GNN w/o G	0.397	0.527	0.361
NTPP-GNN w/o T	0.392	0.525	0.358
NTPP-GNN	0.406	0.538	0.366

表 8 测试集 Gowalla 中擦除实验结果对比

方法	ACC@1	ACC@5	Macro-F1
Bi-TULER	0.252	0.376	0.200
NTPP-GNN w/o G	0.264	0.393	0.211
NTPP-GNN w/o T	0.263	0.392	0.209
NTPP-GNN	0.270	0.404	0.215

表 9 测试集 Brightkite 中擦除实验结果对比

方法	ACC@1	ACC@5	Macro-F1
Bi-TULER	0.556	0.663	0.466
NTPP-GNN w/o G	0.567	0.673	0.476
NTPP-GNN w/o T	0.565	0.670	0.474
NTPP-GNN	0.583	0.681	0.484

总的来说, NTPP-GNN 模型需要依赖于用户时空轨迹涉及的空间、时间、用户社交关系三方面信息作为模型输入, 并且引入这些信息会在一定程度上给计算带来额外开销. 但是, 擦除实验结果说明, 考虑这些信息对于用户轨迹识别任务有所帮助, 具有较为积极意义.

6.3 神经时间点过程方向性影响分析

如前文所述, 关于神经时间点过程的使用往往基于正向时间顺序方向, 即按照事件发生时间的先后顺序计算时间生成概率似然. 由于本文着重于得到序列整体表示用于序列分类而非下一事件发生时间预测, 因此提出双向神经时间点过程对时间序列建模. 为验证点过程方向性对于研究任务性能的影响, 本小节从 NTPP-GNN 模型中去掉社交关系建模部分, 然后针对点过程建模作如下设置:

正向点过程. 将时间序列按照签到时间发生的先后顺序排列, 然后仅使用时间序列建模中的正向部分计算时间生成概率似然.

反向点过程. 与正向点过程类似, 唯一区别在于使用对时间序列建模中的反向部分计算时间生成概率似然.

双向点过程. 使用时间序列模型中的双向部分联合计算时间相关似然.

以上三种设置的主要不同方面在于所构造的时间损失函数. 不同数据集上的实验结果如表 10、表 11 和表 12 所示.

可以看到, 相比单向点过程, 在三个数据集上采用双向点过程对性能有一定提高, 说明所提出双向神经点过程对于用户轨迹识别任务的正面贡献.

表 10 测试集 Foursquare 中不同方向点过程方法对比

方法	ACC@1	ACC@5	Macro-F1
正向点过程	0.395	0.518	0.353
反向点过程	0.393	0.523	0.356
双向点过程	0.397	0.527	0.361

表 11 测试集 Gowalla 中不同方向点过程方法对比

方法	ACC@1	ACC@5	Macro-F1
正向点过程	0.257	0.386	0.207
反向点过程	0.256	0.389	0.204
双向点过程	0.264	0.393	0.211

表 12 测试集 Brightkite 中不同方向点过程方法对比

方法	ACC@1	ACC@5	Macro-F1
正向点过程	0.257	0.386	0.469
反向点过程	0.256	0.389	0.471
双向点过程	0.567	0.673	0.476

6.4 重要参数影响分析

本节主要探讨 NTPP-GNN 模型涉及的一些关键性参数变化对于结果的影响.

(1) 首先分析公式(24)中所示超参 λ 对于实验结果的影响. λ 取值越大代表基于时间生成概率似然的损失在混合优化目标中占比越大, 反之则会削弱其作用. 表 13、表 14 和表 15 展示了不同 λ 取值下三个实验数据集上的结果变化.

从表中结果可以看出, 初始阶段随着 λ 的增大, 三个数据集上多个指标结果均得到一致提升, 这说明刻画时间生成情况有利于用户轨迹识别任务. 但当 λ 增加到一定程度后, 继续加大其取值将造成识

表 13 测试集 Foursquare 中权重系数的影响

权重系数	ACC@1	ACC@5	Macro-F1
0.2	0.397	0.529	0.364
0.4	0.409	0.534	0.369
0.5	0.406	0.538	0.366
0.6	0.404	0.531	0.366
0.8	0.398	0.528	0.368

表 14 测试集 Gowalla 中权重系数的影响

权重系数	ACC@1	ACC@5	Macro-F1
0.2	0.268	0.396	0.206
0.4	0.269	0.400	0.214
0.5	0.270	0.404	0.215
0.6	0.273	0.402	0.218
0.8	0.266	0.400	0.205

表 15 测试集中 Brightkite 上权重系数的影响

权重系数	ACC@1	ACC@5	Macro-F1
0.2	0.574	0.680	0.472
0.4	0.580	0.683	0.480
0.5	0.583	0.681	0.484
0.6	0.585	0.685	0.486
0.8	0.577	0.678	0.478

别性能下降. 这一现象亦符合预期, 因为混合优化目标中的第一部分损失与任务目标直接相关, 过分降低其相对占比将使模型学习偏离目标. 总的来说, 在 Foursquare、Gowalla 和 Brightkite 三个数据集上当 λ 处于 0.4 到 0.6 范围内时, 能够得到较好且稳定的结果.

(2) 接下来分析用户轨迹序列长度对实验结果的影响. 直观上, 若序列长度较短, 则含有的信息量相对匮乏, 可能造成模型识别难度增大. 为此, 本节将待测试轨迹最长长度分别设定为 3、8、15 和 20 进行实验对比, 在 Foursquare 和 Gowalla 两个测试集上的实验结果如图 4 和图 5 所示.

从结果曲线变化趋势可以看出, 当初始阶段轨迹最大长度从 3 变为 8 时, 各项指标均有明显提升.

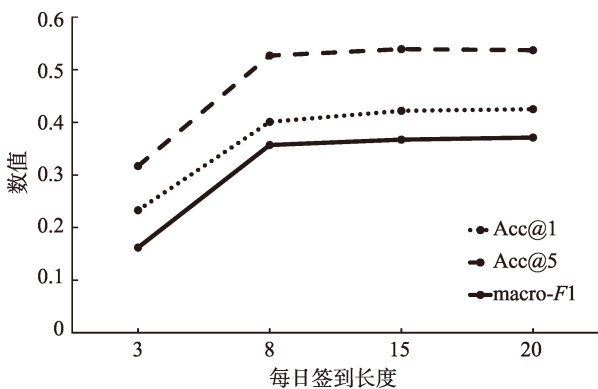


图 4 测试集 Foursquare 轨迹最长长度对实验结果的影响

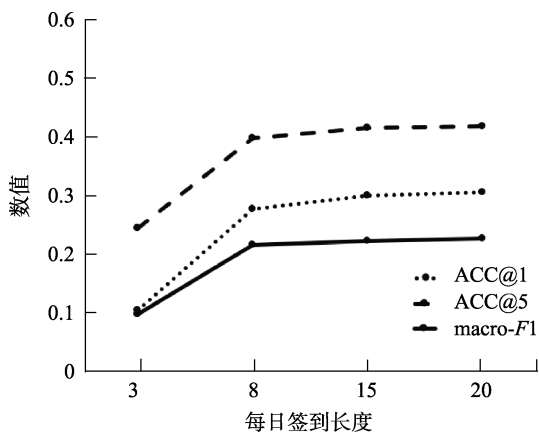


图 5 测试集 Gowalla 轨迹最长长度对实验结果的影响

之后当长度进一步增加时, 各项指标上升幅度趋于减小并逐渐稳定.

(3) 接着进一步分析模型中用户、地点、时间向量化表示维度变化对于用户轨迹识别性能的影响. 为简化建模和分析过程, 默认三种向量维度相等. 具体而言, 本节探索维度取值属于集合 {50, 100, 200, 300, 400} 下的模型结果在 ACC@1 指标上的变化趋势. 如图 6 所示, 三个测试集上的结果均反映出: 在初始阶段, NTPP-GNN 模型性能随着向量维度的增加而提高; 当向量维度超过 200 之后, 模型性能趋于稳定. 上述现象从侧面反映了本文选取的向量维度 300 有其合理性.

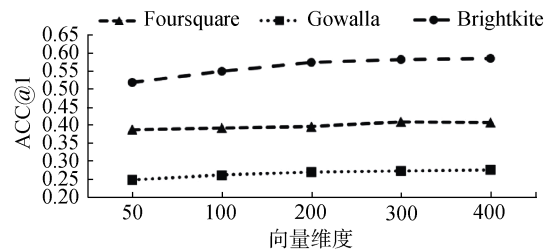


图 6 向量维度变化对 NTPP-GNN 方法效果的影响

(4) 最后, 简要分析模型中注意力头数不同时识别性能如何发生变化. 本节从集合 {1, 2, 3} 中选择注意力头数并进行相应性能测试, 相关结果如表 16 所示. 可以看到, 随着注意力数的增加, 结果有轻微向上增长的趋势.

表 16 注意力头数变化对于结果的影响

数据集	1	2	3
Foursquare	0.405	0.406	0.409
Gowalla	0.266	0.271	0.273
Brightkite	0.577	0.581	0.583

6.5 训练集大小影响分析

本节探索训练集大小变化对于模型性能的影响, 以对比不同方法在具有不同稀疏程度的数据上训练后的效果. 为此, 分别从 Foursquare 和 Brightkite 训练集中选取 100%、80%、60%、40% 和 20% 的数据量训练 NTPP-GNN、BiTULER 和 TULER-LSTM 三种模型. 此时, 对应数据集中用户轨迹数稀疏度也相应发生变化. 例如对于 Brightkite 而言, 对应用户平均轨迹数分别为 10.28, 8.28, 6.33, 4.41, 2.60. Foursquare 和 Brightkite 上关于 ACC@1 指标的测试结果分别如图 7 和图 9 所示, 关于 Macro-F1 指标的测试结果分别如图 8 和图 10 所示.

总的来看, 三种方法随着训练数据量的减少而

不可避免地遭受性能下降问题. 然而, 所提出模型 NTPP-GNN 依然一致地好于性能最好的基准方法 Bi-TULER, 说明模型在具有不同稀疏度的训练数据上训练, 其性能均可以得到一定程度的鲁棒性保证.

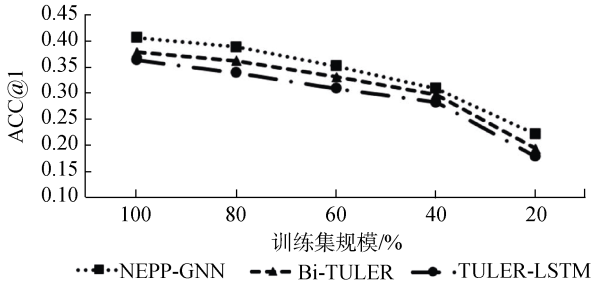


图7 训练集 Foursquare 不同规模大小对 ACC@1 的影响

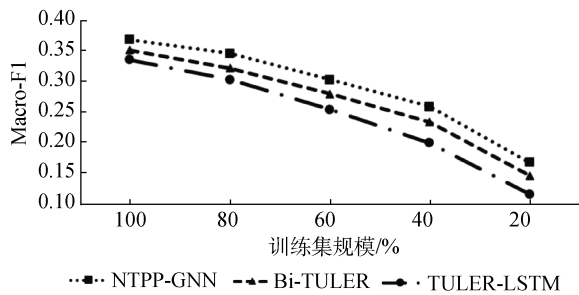


图8 训练集 Foursquare 不同规模大小对 Macro-F1 的影响

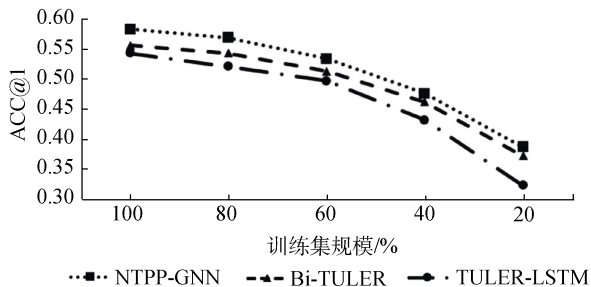


图9 训练集 Brightkite 不同规模大小对 ACC@1 的影响

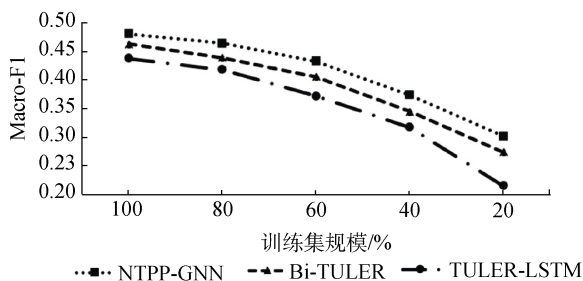


图10 训练集 Brightkite 不同规模大小对 Macro-F1 的影响

6.6 测试集轨迹长度影响分析

为进一步说明方法在轨迹本身具有不同稀疏度时的效果, 本节根据不同轨迹长度区间将测试集划分为不同部分, 并对比相应部分测试数据效果. 具

体而言, 按照长度区间 1 到 5 (1-5)、6 到 10 (6-10)、11 到无穷 (11-Inf) 进行划分, 结果如表 17 和 18 所示.

表 17 测试集 Foursquare 中不同序列长度性能对比

测试轨迹 序列长度	TULER-LSTM		Bi-TULER		NTPP-GNN	
	ACC@1	Macro-F1	ACC@1	Macro-F1	ACC@1	Macro-F1
1-5	0.330	0.289	0.362	0.312	0.388	0.338
6-10	0.368	0.338	0.385	0.354	0.410	0.369
11-Inf	0.391	0.356	0.400	0.372	0.425	0.393

表 18 测试集 Brightkite 中不同序列长度性能对比

测试轨迹 序列长度	TULER-LSTM		Bi-TULER		NTPP-GNN	
	ACC@1	Macro-F1	ACC@1	Macro-F1	ACC@1	Macro-F1
1-5	0.515	0.415	0.532	0.449	0.557	0.473
6-10	0.556	0.463	0.578	0.481	0.604	0.499
11-Inf	0.588	0.487	0.611	0.494	0.631	0.523

整体而言, 在稀疏的短轨迹数据上, 不同方法的效果相对较差. 与之相对的是, 随着序列长度的增加, 各方法所能利用的信息量也有一定的增长, 因此性能均取得一定程度的提升. 横向对比各个方法可以发现, NTPP-GNN 在短轨迹和长轨迹数据上效果均优于 Bi-TULER 和 TULER-LSTM, 说明 NTPP-GNN 在不同长度的测试轨迹上用户识别性能具有较好的鲁棒性.

7 总结与展望

本文研究了用户轨迹识别问题, 以判断目标轨迹所属用户, 对于多项下游应用均有促进作用. 本文强调了该研究问题面临的用户轨迹稀疏和类别数量庞大这两项关键性挑战, 并尝试从利用轨迹所含时间序列信息和用户之间社交关系两个角度应对上述挑战. 为此, 本文提出了融合神经时间过程与图神经网络的模型 NTPP-GNN. 该模型在轨迹地点序列建模的基础上, 一方面提出双向神经时间过程学习时间序列表征, 丰富轨迹序列表示; 另一方面通过图神经网络刻画社交关系网络中蕴含的用户关联关系, 强化用户表示. NTPP-GNN 采用端到端学习方式统一不同建模部分以使其彼此适配. 在三组公开数据集上的实验结果验证了 NTPP-GNN 达到了当前最佳性能, 并揭示了时间序列建模和社交关系建模对于研究问题的正面贡献.

未来主要考虑从以下三方面拓展当前工作: (1) 目前方法还未能针对新用户 (测试集中用户在训练集中未出现) 进行特别处理. 计划借助零样本/少样

本学习技术^[48], 通过有效利用用户的一些基本属性信息实现新用户轨迹识别; (2) 随着包含位置信息社交网络的普及, 用户广泛存在于不同平台中, 因此如何利用一个平台中的用户轨迹信息帮助另一个平台中的用户轨迹识别问题也是值得开展的一项研究工作^[49]; (3) 如何在充分保护用户隐私的条件下进行用户轨迹识别。

参 考 文 献

- [1] Zheng Y. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology*, 2015, 6(3): 1-41
- [2] Yin H Z, Cui B. *Spatio-temporal recommendation in social media*. New York, USA: Springer, 2016
- [3] Mao Jia-Li, Jin Che-Qing, Zhang Zhi-Gang, Zhou Ao-Ying. Anomaly detection for trajectory big data: advancements and framework. *Journal of Software*, 2017, 28(1): 17-34 (in Chinese) (毛嘉莉, 金澈清, 章志刚, 周傲英. 轨迹大数据异常检测: 研究进展及系统框架. *软件学报*, 2017, 28(1): 17-34)
- [4] Zhou X, Noulas A, Mascolo C, Zhao Z X. Discovering latent patterns of urban cultural interactions in WeChat for modern city planning//*Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining*. London, UK, 2018: 1069-1078
- [5] Liu Q, Wu S, Wang L, Tan T N. Predicting the next location: a recurrent model with spatial and contexts//*Proceedings of the 30th Conference on Artificial Intelligence*. Phoenix, Arizona, USA, 2016: 194-200
- [6] Feng J, Li Y, Zhang C, Sun F N, Meng F C, Guo A, Jin D P. DeepMove: predicting human mobility with attentional recurrent networks//*Proceedings of the 2018 World Wide Web Conference on World Wide Web*. Lyon, France, 2018: 1459-1468
- [7] Zhang W, Wang J Y. Location and time aware social collaborative retrieval for new successive point-of-interest recommendation//*Proceedings of the 24th International Conference on Information and Knowledge Management*. Melbourne, Australia, 2015: 1221-1230
- [8] Hidasi B, Karatzoglou A, Baltrunas L, Tikk D. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv: 1511.06939*, 2015
- [9] Quadana M, Karatzoglou A, Hidasi B, Cremonesi P. Personalizing session-based recommendations with hierarchical recurrent neural networks//*Proceedings of the 11th Conference on Recommender Systems*. Como, Italy, 2017: 130-137
- [10] Song C M, Qu Z H, Blumm N, Barabasi A L. Limits of predictability in human mobility. *Science*, 2010, 327(5968): 1018-1021
- [11] Xu F L, Tu Z, Li Y, Zhang P Y, Fu X M, Jin D P. Trajectory recovery from ash: user privacy is not preserved in aggregated mobility data//*Proceedings of the 26th International Conference on World Wide Web*. Perth, Australia, 2017: 1241-1250
- [12] Kurashima T, Iwata T, Hoshida T, Takaya N, Fujimura K. Geo topic model: joint modeling of user's activity area and interests for location recommendation//*Proceedings of the 6th International Conference on Web Search and Data Mining*. Rome, Italy, 2013: 375-384
- [13] Yin H Z, Sun Y Z, Cui B, Hu Z T, Chen L. LCARS: a location-content-aware recommender system//*Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*. Chicago, USA, 2013: 221-229
- [14] Zheng V W, Zheng Y, Xie X, Yang Q. Collaborative location and activity recommendations with GPS history data//*Proceedings of the 19th International Conference on World Wide Web*. Raleigh, USA, 2010: 1029-1038
- [15] Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation//*Proceedings of the 19th International Conference on World Wide Web*. Raleigh, USA, 2010: 811-820
- [16] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*, 1997, 9(8): 1735-1780
- [17] Kalchbrenner N, Blunsom P. Recurrent continuous translation models// *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, USA, 2013: 1700-1709
- [18] Gao Q, Zhou F, Zhang K, Trajcevski G, Luo X, Zhang F L. Identifying human mobility via trajectory embeddings//*Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Melbourne, Australia, 2017: 1689-1695
- [19] Zhou F, Gao Q, Trajcevski G, Zhang K P, Zhong T, Zhang F L. Trajectory-user linking via variational autoencoder//*Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 2018: 3212-3218
- [20] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. *Nature*, 2006, 439: 462-465
- [21] Tang J, Qu M, Wang M Z, Zhang M, Yan J, Mei Q Z. LINE: large-scale information network embedding//*Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy, 2015: 1067-1077
- [22] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 2007, 58(7): 1019-1031
- [23] Hawkes A. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971, 58(1): 83-90
- [24] Du N, Dai H J, Trivedi R, Upadhyay U, Gomez-Rodriguez M, Song L. Recurrent marked temporal point processes: embedding event history to vector//*Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016: 1555-1564
- [25] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017
- [26] Monreale A, Pinelli F, Trasarti R, Giannotti F. WhereNext: a location predictor on trajectory pattern mining//*Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining*. Paris, France, 2009: 637-646
- [27] Cheng C, Yang H Q, Lyu M R, King I. Where you like to go next: successive point-of-interest recommendation//*Proceedings of the 23rd International Joint Conference Artificial Intelligence*. Beijing, China, 2013: 2605-2611
- [28] Qin J R, Ren K, Fang Y C, Zhang W N, Yu Y. Sequential recommendation with dual side neighbor-based collaborative relation modeling//*Proceedings of the 13th ACM International*

- Conference on Web Search and Data Mining. Houston, USA, 2020: 465-473
- [29] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality// Proceedings of the 27th Annual on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111-3119
- [30] Kingma D, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv: 1312. 6114, 2013
- [31] Ogata Y. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 1998, 50(2): 379-402
- [32] Gao S, Ma J, Chen Z M. Modeling and predicting retweeting dynamics on microblogging platforms//Proceedings of the 8th International Conference on Web Search and Data Mining. Shanghai, China, 2015: 107-116
- [33] Mei H, Eisner J. The neural hawkes process: a neurally self-modulating multivariate point process//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6754-6764
- [34] Okawa M, Iwata T, Kurashima T, Tanaka Y, Toda H, Ueda N. Deep mixture point processes: spatio-temporal event prediction with rich contextual information//Proceedings of the 25th International Conference on Discovery and Data Mining. Anchorage, USA, 2019: 373-383
- [35] Liang W, Zhang W, Wang X L. Deep sequential multi-task modeling for next check-in time and location prediction// Proceedings of the 24th International Conference on Database Systems for Advanced Applications. Chiang Mai, Thailand, 2019: 353-357
- [36] Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016
- [37] Wu Z H, Pan S R, Chen F W, Long G D, Zhang C Q, Yu P. A Comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020
- [38] Yang C, Xiao Y, Zhang Y, Sun Y, Han J. Heterogeneous network representation learning: survey, benchmark, evaluation, and beyond. arXiv preprint arXiv:2004.00216, 2020
- [39] Yu B, Yin H T, Zhu Z X. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 3634-3640
- [40] Bai L, Yao L, Kanhere S, Wang X Z, Shen Q Z. STG2Seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 1981-1987
- [41] Ma Y, Wang Y Q, Jin W, Wang X, Tang J L, Jia C Y, Yu J. Traffic flow prediction via spatial temporal graph neural network// Proceedings of the Web Conference 2020. Taipei, China, 2020: 1082-1092
- [42] Wang X, He X N, Feng F L, Chua T S. Neural graph collaborative Filtering//Proceedings of the 42nd International Conference on Research and Development in Information Retrieval. Paris, France, 2019: 165-174
- [43] Wang W, Zhang W, Liu S K, Liu Q, Zhang B, Lin L Y, Zha H Y. Beyond clicks: modeling multi-relational item graph for session-based target behavior//Proceedings of the Web Conference 2020. Taipei, China, 2020: 3056-3062
- [44] Jin D, Liu Z Y, Li W H, He D X, Zhang W X. Graph convolutional networks meet markov random fields: semi-supervised community detection in attribute networks// Proceedings of the 33rd Conference on Artificial Intelligence. Honolulu, USA, 2019: 152-159
- [45] Yang D Q, Qu B Q, Yang J, Cudre-Maurou P. Revisiting user mobility and social relationships in LBSNs: a hypergraph approach//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 2147-2157
- [46] Cho E, Myers S, Leskovec J. Friendship and mobility: user movement in location-based social networks//Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1082-1090
- [47] Kingma D, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014
- [48] Satorras V, Estrach J. Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043, 2017
- [49] Feng J, Zhang M Y, Wang H D, Yang Z Y, Zhang C, Li Y, Jin D P. Dplink: user identity linkage via deep neural network from heterogeneous mobility data//Proceedings of the World Wide Web Conference 2019. San Francisco, USA, 2019: 459-469



ZHANG Wei, Ph.D., associate professor. His main research interest is data mining.

LI Yang, M. S. His main research interest is personalized user modeling.

ZHANG Ji, Ph.D., professor. His main interests include big data analysis, data mining, and information privacy protection and security.

WANG Jian-Yong, Ph.D., professor. His main research interest is data mining.

Background

The studied user trajectory identification problem belongs to user sequential behavior modeling in the fields recommender system and user data mining. Current efforts

in this regard are devoted to designing effective sequential modeling approaches to learn user representations and capture user long-term and short-term interests. Sequential

neural modeling approaches, such as RNNs, Transformer networks, and CNNs, have been demonstrated to be effective in different tasks belonging to user sequential behavior modeling. However, only a few studies have investigated the critical problem of user trajectory identification. They rely on bi-directional RNNs to learn trajectory representations by only regarding location sequences as input. The learned representations are then fed into fully-connected networks for inferring the possible users who generate these trajectories.

This paper addresses the two main challenges faced by the problem: trajectory sparsity and large-scale classes. To handle these challenges, this paper proposes to leverage temporal information of trajectories and social relations between users to enhance the representations trajectories and users. A novel neural model is developed by coupling

neural temporal point process with graph neural network. Thus, it could incorporate the two types of information seamlessly. The extensive experiments demonstrate that the proposed model could improve the current state-of-the-art approach on two publicly available datasets. To be specific, ACC@5 is improved from 0.512 to 0.538 on the Foursquare dataset and from 0.376 to 0.404 on the Gowalla dataset.

The authors of this paper already have done some studies relevant to user trajectory modeling(such as successive POI recommendation based on user trajectory modeling(please refer to the work^[7]) and next check-in time and location prediction(please refer to the work^[35])) and user generated content modeling in a broader aspect.

This work is supported by National Natural Science Foundation of China (62072182, 61702190, 61532010) and Zhejiang Lab (2019KB0AB04).