

# 位置社交网络上的图表示学习

赵琳琳<sup>1)</sup> 吴安彪<sup>1)</sup> 袁 野<sup>2)</sup> 李 扬<sup>1)</sup> 王国仁<sup>2)</sup>

<sup>1)</sup> (东北大学计算机科学与工程学院 沈阳 110167)

<sup>2)</sup> (北京理工大学计算机学院 北京 100081)

**摘 要** 随着在线社交网络的普及,基于位置的社交网络(Location-Based Social Networks, LBSN)积累了海量数据,且因其包含丰富的时空、语义信息已被广泛应用在挖掘用户行为偏好的研究上,但传统的手工提取LBSN特征的方法有很大局限性且耗时耗力.近几年来,图表示学习在推荐系统、知识图谱等领域成功应用,彰显了其强大的非线性拟合和表示学习的能力.然而,现有图表示学习大多集中在静态、同构的网络上,难以同时考虑时间、位置信息、社交关系来捕捉LBSN中复杂的结构和用户偏好,以致无法高效提取LBSN中的有效信息.因此,本文提出面向LBSN的两阶段图表示学习框架TGE-LBSN (Two Stages of Graph Embedding on LBSN),即将LBSN转化成异构网络结构,设计了LBSN上的图表示学习算法自动提取LBSN的特征,得到蕴含有效信息的节点向量表示,并利用社交领域的预测、推荐任务检验其有效性.首先,依据时间对LBSN的签到(Check-in)超边进行有偏采样,第一阶段设计了IVGS (Initial Vector Generation Stage)算法,利用好友边与Check-in超边共同生成包含位置、特征信息的初始节点向量.其次,在第二阶段将LBSN依据签到时间划分成不同子图,分别进行各个子图下的异构网络层结构信息聚合操作.在第一阶段结果的基础上,提出了面向LBSN的选择聚合邻居策略SAN (Select Aggregated Neighbors),选取有代表性的邻居节点完成聚合操作,进而完成子图向量生成算法SVG (Subgraph Vector Generation)得到子图中节点的向量表示.最后,依据任务设定损失函数,结合注意力机制为各子图学得自适应权重,从而得到节点的最终向量表示,进而完成社交领域的预测推荐任务.本研究分别在真实的LBSN数据集上以及时序社交网络与基准方法进行了大量的对比实验,并采用ROC曲线作为评价标准,实验结果验证了本文所提算法TGE-LBSN能高效地自动提取LBSN的有效信息并保留在节点的嵌入向量中,且在社交领域的好友预测任务上比现有模型在AUC值方面最高可提升42%,在兴趣点推荐任务上AUC取值相较于对比算法最高可达到7%的提升.

**关键词** 位置社交网络;图表达;异构网络表示学习;注意力机制;链路预测

中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2022.00838

## Graph representation learning on Location-Based Social Networks

ZHAO Lin-Lin<sup>1)</sup> WU An-Biao<sup>1)</sup> YUAN Ye<sup>2)</sup> LI Yang<sup>1)</sup> WANG Guo-Ren<sup>2)</sup>

<sup>1)</sup> (School of Computer Science and Engineering, Northeastern University, Shenyang 110167)

<sup>2)</sup> (School of Computer, Beijing Institute of Technology, Beijing 100081)

**Abstract** With the popularity of online Social Networks, location-based social networks (LBSN) have accumulated massive data and have been widely used in the research of mining user behavior preferences due to their rich Spatio-Temporal and semantic information. Nevertheless, the traditional manual extraction of LBSN features is limited and time-consuming. In recent years, graph representation learning has been successfully applied to the modeling and representation of

收稿日期:2021-07-12;在线发布日期:2021-12-01. 本课题得到国家自然科学基金(61932004, 61732003, 61729201, 61972077)、辽宁振兴人才计划(No. XLYC2007079)资助. 赵琳琳, 硕士研究生, 中国计算机学会(CCF)学生会会员, 主要研究领域为图神经网络、位置社交网络. E-mail: zll997910@foxmail.com. 吴安彪, 博士研究生, 中国计算机学会(CCF)学生会会员, 主要研究领域为图数据库、图神经网络. 袁 野(通信作者), 博士, 教授, 国家优秀青年基金获得者, 中国计算机学会(CCF)高级会员, 主要研究领域为云计算、大数据管理(包括图数据管理、不确定数据管理、数据隐私保护)、P2P计算等. E-mail: yuan-ye@bit.edu.cn. 李 扬, 硕士研究生, 中国计算机学会(CCF)学生会会员, 主要研究领域为图嵌入、图神经网络. 王国仁, 博士, 教授, 长江学者、国家杰出青年科学基金获得者, 中国计算机学会(CCF)会员, 主要研究领域为不确定图管理、数据密集型计算、非结构化数据管理、分布式查询处理和优化技术.

various graph structure data such as recommendation systems and knowledge maps, demonstrating its powerful non-linear fitting and representation learning capabilities. However, most of the existing Graph representation learning studies focus on static and homogeneous networks, and it is difficult to simultaneously combine time, location information, and social relationships to capture the complex structure and user preferences in LBSN, which makes it difficult to extract effective information from LBSN. Therefore, this paper proposes a two-stage Graph representation learning framework TGE-LBSN (Two Stages of Graph Embedding on LBSN) for LBSN, which transforms LBSN into a heterogeneous network, and automatically extracts the features of LBSN with the help of graph representation learning to obtain nodes' vector representation with sufficiently rich information and utilize the prediction and recommendation tasks in the social domain to verify its effectiveness. First of all, biased sampling is carried out on the check-in hyperedge of LBSN according to users' Check-in time. In the first stage, the IVGS (Initial Vector Generation Stage) algorithm is designed, the friendship edges and the Check-in super edges are used to jointly generate nodes' vectors containing position and feature information by IVGS algorithm. The generated nodes' vectors are used as the input of the second stage. In addition, the second stage is mainly responsible for generating the final nodes' vectors in LBSN. LBSN is divided into different subgraphs according to the Check-in time, and we design the LBSN-oriented SAN (Select Aggregated Neighbors) strategy which is used to select representative neighbors to complete the aggregation operation, and then use the subgraph vector generation algorithm SVG (Subgraph Vector Generation) to obtain the vector representation of the nodes in each subgraph. Finally, the loss function is set according to the downstream tasks, and the attention mechanism is also used to learn adaptive weights for the subgraphs in different time periods to obtain the final vectors of the nodes, and then we use the final nodes' vectors to complete various prediction tasks in the social domain. Plenty of comparative experiments are carried out with the benchmark methods on the real LBSN data sets and on the time series social network, respectively, we use ROC curve as the evaluation standard. Extensive experimental results verify that the proposed algorithm TGE-LBSN outperforms other benchmark methods, and it can efficiently extract the effective information of LBSN and retain it in the embedding vector of the node. Specifically, in terms of friendship prediction, the AUC value can be increased by up to 42% compared with the existing models. On the point of interest recommendation task, the AUC value can be increased by up to 7% compared with the benchmark algorithm.

**Keywords** graph embedding; heterogeneous network representation learning; attention mechanism; link prediction

## 1 引言

随着智能终端设备的普及,使得获取和处理位置信息变得方便快捷,越来越多人通过移动终端使用在线社交网络,同时积累了海量的基于位置的社交网络数据,这也为挖掘其中隐藏的有价值信息提供了机遇与挑战.经典的LBSN数据集如Foursquare、Gowalla、Yelp、BrightKite等等,其中Foursquare自

2009年成立以来,已经拥有用户1000万,日签到次数达到300万次.目前对异构时空网络特别是LBSN上的社交用户轨迹信息进行数据挖掘已成为新兴的研究热点,相较于传统社交网络,LBSN除了用户之间社交关系之外,增加了多元化的数据,比如用户Check-in信息、地理位置信息、活动点语义信息等. LBSN数据形式如图1所示,整个LBSN大致包括两层网络,社交关系层和签到异构层,其中实线代表用户之间的社交关系、友谊关系组成了社交网络层;

Check-in边是指移动设备上系统自动实时获取用户位置信息,其中位置、签到时间信息与用户、活动节点共同构成签到异构网络层,如图中虚线所示.用户通过移动设备以签到形式记录自己的位置空间信息,并分享对兴趣点(Point Of Interest, POI),如书店、电影院、图书馆、餐厅等的评价及体验,即Check-in数据表述了某个用户在某个时间打卡某地并做出评价.位置与时间的加入使得社交网络所蕴含的信息更加丰富,为社交领域的预测分析提供了更多可能.

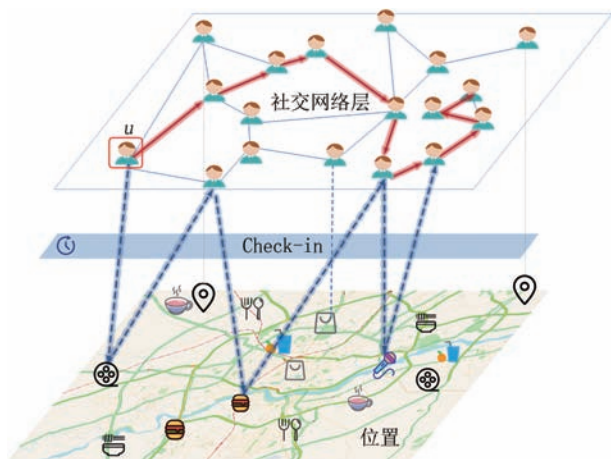


图1 LBSN示意图

传统针对LBSN的研究大部分集中在对LBSN数据进行集中建模来分析用户偏好,比如矩阵分解、对不同位置的兴趣爱好进行分析建模、评估打卡过的兴趣点的地理相关性作为距离分布、协同过滤法等等<sup>[1-5]</sup>,或是利用社交网络建模分析用户偏好从而进行商品推荐<sup>[6]</sup>.但传统方法具有很大局限性,且仅针对社交领域单一任务进行建模分析,可扩展性差,甚至可能需要矩阵分解,计算复杂度高,且预测推荐效果并不理想.

近年来在社交网络、生物信息、化学等领域中大量采用图结构作为数据表示,同时图表示学习也逐渐成为新兴的研究方向,越来越多学者借助图表示学习解决图结构的数据研究问题.文献[7]借助图表示学习强大的表达能力,将LBSN与DeepWalk<sup>[8]</sup>相结合,来获取LBSN签到数据中隐含的丰富信息,且学得向量可以用作社交领域各类预测推荐任务,但未能充分利用LBSN社交网络层的友谊信息,且无法聚合用户和兴趣点自身的特征信息.所以,将基于游走的图表示学习与图神经网络相结合应用在挖掘LBSN潜在有价值信息上.

LBSN蕴含着用户大量的轨迹信息,用户行为轨迹形成了LBSN关系多样化、动态演变化、错综复杂化等特点,同时为用户行为偏好、影响力分析、推荐系统等领域的研究提供了更多可能.但由于LBSN数据稀疏、结构复杂、属性多样化以及学习任务多层化,使得丰富的上下文信息以及复杂结构不易被挖掘.无论是针对同构网络还是异构网络的表示学习都很难高效获取LBSN中隐藏的潜在上下文语义信息及复杂的结构信息,同时结构信息包含的拓扑结构对于图嵌入生成良好的表示至关重要<sup>[9]</sup>.所以针对LBSN的研究面临着如下的挑战:(1)如何有效地保留LBSN中多类型节点之间上下文语义信息?(2)如何充分利用异构时空网络特有的时空信息?在LBSN上是否可以直接利用异构网络的表示学习框架?(3)如何衡量LBSN中各类节点之间的相似度?(4)如何在向量投影空间体现用户与兴趣点之间的签到关系?

为了应对以上挑战,本研究为LBSN设计了特定的游走策略以及邻居聚合算法,以保留多类型节点的上下文语义信息.同时为了充分利用时空信息提出了有偏采样以及时序划分策略.本研究针对LBSN设计一种图表示学习方法,主要目的在于充分挖掘LBSN的有价值信息,将LBSN的高维网络结构降低到低维空间,并利用神经网络进行自动特征提取,依据设计的损失函数,确保LBSN原始网络的某些性质在向量空间中也能得到对应,使得在向量空间中保持用户节点和兴趣点的签到关系以及用户间的好友关系,将用户签到行为偏好信息保留在节点特征向量表示中.为此,本文提出了一种基于异构时空网络LBSN的图表示学习框架TGE-LBSN,对LBSN丰富的上下文信息进行分阶段提取,目标是在向量投影空间保留LBSN中的有效信息,再依据签到时间进行多时间段划分,并设计了选择聚合邻居策略,最终得到当前子图下的节点向量表示.最后,依据预测任务设定损失函数,借助注意力机制学得各个时间段的自适应权重,将各子图下的特征向量进行加权汇合,从而利用最终节点的向量表示完成社交领域的各种推荐预测任务以进行有效性验证,如好友推荐、兴趣点推荐、用户行为轨迹预测等.本文主要创新点如下:

(1) 本文针对LBSN表示学习面临的独特挑战,设计出一种新型的TGE-LBSN算法,对其进行自动特征提取,以实现在LBSN上的图表示学习,并给出形式化异构时空网络LBSN表示学习问题定义;



(2) 面对 LBSN 网络中存在签到超边的情况, 首先设计一种根据 Check-in 签入时间以及节点重要程度的有偏采样策略, 保证采样策略得到的数据能尽可能与原始数据中超边数保持一致的分布;

(3) 本文设计两阶段信息聚合学习算法 TGE-LBSN, 阶段一将采样之后的数据集, 设计特有的两层游走策略, 利用 IVGS 算法获取初始节点向量. 阶段二依据签到时间进行子图划分, 并设计了一种选择邻居聚合 SAN 策略, 再利用 SVG 算法将签到网络层结构信息保留在节点特征向量中, 获得子图下的节点向量; 进一步借助注意力机制, 为不同时间段子图的表示向量学得自适应的权重, 以融合各子图节点向量得到最终向量表示;

(4) 通过在真实数据集上的实验, 验证了 TGE-LBSN 算法学得节点向量表示能有效地聚合 LBSN 上的复杂结构信息以及上下文语义信息, 并在好友预测、兴趣点推荐的任务上表现出不错的效果.

本文第 2 节对 LBSN 及图表示学习的近期研究工作做出简单介绍; 第 3 节给出基本概念及问题定义; 第 4 节对 LBSN 进行预处理, 以及生成初始的向量; 第 5 节详细介绍了 TGE-LBSN 框架, 以及聚合邻居选择策略和最终向量表示的生成过程; 第 6 节在真实数据集上对算法进行实验和结果分析; 第 7 节主要对全文的总结以及对未来的展望.

## 2 相关工作

本章节主要对现阶段关于 LBSN 的相关研究、图表示学习算法相关介绍以及目前的进展、LBSN 与图表示学习相结合的算法进行了简单介绍, 为之后介绍 TGE-LBSN 算法做相关铺垫.

### 2.1 LBSN 相关研究

在过去的几年中, 基于位置的社交网络已被广泛用于挖掘社交领域用户偏好的研究上. 现阶段针对 LBSN 的研究大部分集中在对 LBSN 进行建模, 文献[1]提出一种矩阵分解模型, 借助地理相关性, 生成地理相关分数融进概率矩阵分解模型. 文献[2]通过对用户的阅读的地理位置进行聚类, 从而分析出用户在不同位置的阅读偏好. 文献[3]提出了一种双重社区聚类与关联方法, 通过信息熵度量用户发布主题之间的相似性, 采用超边聚类方式获取用户关于地理位置的兴趣点特征簇以得到局部最优解. 文献[4]认为用户的活动空间服从幂律分布, 借助朴素贝叶斯, 提出协同推荐算法. 文献[5]通过

用户签到的位置聚集现象嵌入到用户和 POI 的隐向量中, 建模用户签到的地理影响. 文献[10]将时间循环模式整合到了基于用户的协作过滤框架中, 以进行时间感知的 POI 推荐.

传统的针对 LBSN 研究方法都旨在为某一个特定的任务进行建模分析, 如文献[2-3, 5, 10]仅针对兴趣点推荐任务, 无法再延伸到其他任务上, 可扩展性差. 这类方法大多计算复杂度高、推荐效果并不显著, 在提取 LBSN 特征信息上不仅需要人工的繁琐工作, 而且很难一概而论所以借助人工作定义的特征工程由于对特征依赖度极高具有很大局限性.

### 2.2 图表示学习

随着图表示学习的快速发展, 越来越多学者倾向于将图表示学习应用在图数据建模领域, 旨在将图结构数据映射到低维的向量空间, 得到图数据的向量表示, 图数据的表示可以是节点层面、边层面, 甚至整个图层面<sup>[11]</sup>, 这些图数据的表示为对应领域更好地利用机器学习提供了桥梁, 同时为图表示学习带来了广泛的应用场景, 例如借助图表示学习进行相关推荐<sup>[12-13]</sup>. 相较于传统的 LBSN 个性化推荐算法, 图表示学习的优势在于无须对用户节点、兴趣点之间的相似性进行建模分析来手动提取 LBSN 中的特征, 通过游走或者聚集转换邻域内的节点特征来提取有效的节点嵌入, 同时捕获各种模式中复杂的内部关系<sup>[14]</sup>. 图表示学习的目的就在于利用神经网络学得包含原始网络中丰富信息的向量表示, 作为下游相关任务的输入, 从而在节点分类、链路预测、图分类等任务实现更好的效果. 图表示学习根据图中边的种类可以划分为两种, 分别是同构网络表示学习和异构网络表示学习, 接下来将详细介绍这两类表示学习的相关方法及经典算法.

#### 2.2.1 同构网络表示学习

现有经典的图表示学习是针对同构图<sup>[8, 10, 15-19]</sup>开发的, 即节点及边属于单一类别. 目前这类图表示学习从方法上来说主要包括三种: 基于分解的方法、基于游走的表示学习、基于深度学习的 GNN. 基于分解的方法通过对图数据的结构信息矩阵进行矩阵分解操作<sup>[15-16]</sup>, 但结果依赖矩阵分解计算, 计算复杂度高. 基于游走的表示学习主要有 DeepWalk<sup>[8]</sup>, LINE<sup>[17]</sup> 和 node2vec<sup>[18]</sup> 等等. 基于深度学习的图神经网络主要有基于频谱的方法和基于空间的方法<sup>[20-23]</sup>. GCN<sup>[20]</sup> 是一种简化的频谱方法, 通过整合当前节点与其邻居节点的特征信息完成对节点属性信息和结构信息的共同学习, 文献[21]证明了

GCN倾向于保留结构信息,而无法保留节点的特征相似度.之后提出了更多基于空间的方法具有不同设计的基于空间的方法来聚合和转换邻域信息,包括GraphSAGE<sup>[22]</sup>、GAT<sup>[23]</sup>等.这类算法本质上都是在GCN基础上对邻居聚合操作做出了改进,GraphSAGE<sup>[22]</sup>仅仅需要考虑k阶邻居,无须考虑全图结构信息,且适合做归纳学习;GAT<sup>[23]</sup>在一定程度上可以提高了图神经网络的表达能力.

然而,大多数社会信息网络本质上是异构网络,涉及多样性的节点和边的类型<sup>[24]</sup>.无属性同构图的表示学习无法从签入超边缘捕获信息,即使通过将每个超边分解为多个经典边缘,将超图转换为经典图,但这种不可逆的过程必然会损失一定的信息,无法很好地表示异构时空网络数据,进而导致学习的节点嵌入效果不佳.

### 2.2.2 异构网络表示学习

现阶段很多学者致力于解决异构网络的表示学习,文献[25]提出了异构网络的嵌入表示,可以捕获不同视图中节点嵌入之间的复杂关系,以及保留邻近信息,但是并不能保留用户及兴趣点的自身的特征信息. Metapath2Vec<sup>[26]</sup>基于给定的元路径的针对异构网络进行游走以获取结构信息.针对异构网络设计图嵌入方法如Hin2VEC<sup>[27]</sup>、HEBE<sup>[28]</sup>、DHNE<sup>[29]</sup>、HGT<sup>[30]</sup>等等,针对不同的异构网络设计了相应的方式,利用深度学习或者原路径改进得到异构网络的节点或边的嵌入表示.文献[31]将四个关系二部图的嵌入编码到一个共享的低维空间中,通过动态学习用户的最新偏好来实现实时POI推荐. LBSN2vec++<sup>[7]</sup>针对LBSN提出基于随机游走的策略联合采样用户签入和社交关系,自动特征学习的异构超图嵌入方法,并从采样的超边缘中学习节点嵌入,但并未考虑社交网络层次的社交用户之间的友谊,且仅在单一数据集<sup>[32]</sup>上进行测试.对于用户节点来说,友谊信息和签到信息都对用户节点的向量生成有很大的影响力.在针对LBSN进行表示学习过程中,节点自身初始特征向量、结构信息、上下文语义信息等均对节点向量生成至关重要.

在时空网络的图表示学习方面,文献[33]首次将图卷积应用到交通预测问题上,对时空网络的分析采用STGCN框架,选择速度、流量、密度来显示监视当前的交通状态,捕获时间、空间关联,用于实时交通预测.文献[34]利用GCN学习空间信息,利用TCN、Gated-TCN组件学习时间信息,并利用GCN处理高阶时空信息.但由于文献[33-34]侧重

于实时交通预测,对于时效性、空间性的要求高,本质上利用节点的时空属性信息,而LBSN是异质超图,节点包含多种类型,而且两个领域的预测任务对于时间的要求也不同,所以并不适合社交领域的时空网络.同时GCN也被拓展应用在知识图谱领域,如R-GCN<sup>[35]</sup>为处理具有高度关系数据特征的现实知识库,将知识图谱的三元组作为输入,应用在链路预测(恢复缺失的元组)和实体分类(为实体分配类型或者分类属性),由于其主要被应用在知识图谱上,与LBSN的结构并不相似,所以无法被直接应用在LBSN上.

由于LBSN拥有丰富而复杂的数据,且编码着复杂的用户社交的动态行为轨迹,借助这些异构网络表示学习算法并不能完全捕获LBSN异构时空网络中的丰富的语义信息,所以针对LBSN这类异构时空的数据需要设定针对性的图嵌入方法,以便更好地挖掘LBSN中隐藏的丰富信息.现有LBSN的表示学习方法对于超图的异质性、复杂数据结构和特征可能无法完全捕获,且未考虑节点自身特征向量.基于此本文设计一种面向LBSN异构超图的嵌入学习方法TGE-LBSN,旨在得到能有效表示LBSN节点的向量表示,并通过完成社交领域的各类预测任务来验证TGE-LBSN算法搭建的表示学习模型获得的节点向量是否包含原始LBSN中足够丰富的信息.

## 3 基本概念与问题定义

本章节主要对基本的概念及问题定义进行梳理,以及研究对象LBSN概念相关介绍,并给出了针对LBSN的异构时空网络的图表示学习的形式化定义及本研究的问题定义.表1是对本文所常用的符号表示及其意义进行汇总并作以简要说明.

### 3.1 基本定义

**定义 1.** LBSN. 给定一个网络  $G=(V,E)$ ,  $V:\{U,A,T,L\}$ ,  $E:\{e,c\}$  表示 LBSN 异构时空网络,  $U$  表示用户节点  $u$  集合,  $A$  代表兴趣点  $a$  的类别集合,  $T$  代表时间节点  $t$  集合,  $L$  代表为位置节点  $l$  集合, 用户节点之间的好友边  $e$  表示为  $(u_m, u_n)$ , 其中  $u_m, u_n \in U$ ,  $c$  代表 Check-in 数据, 表示为  $(u, a, t, l)$ ,  $u \in U, a \in A, t \in T, l \in L$ .

以图1为例, LBSN由社交网路层和异构签到层组成. 好友边  $e$  如图1中实线所示, 好友边  $e$  形成了社

表1 常用符号表示

| 名称          | 描述            | 名称             | 描述           |
|-------------|---------------|----------------|--------------|
| $V$         | LBSN节点集合      | $T$            | 签到时间节点集合     |
| $G$         | LBSN异构网络图     | $U$            | 用户节点集合       |
| $G_S$       | LBSN社交网络层     | $L$            | 签到位置节点集合     |
| $G_A$       | LBSN异构签到网络    | $A$            | 活动类别集合       |
| $e$         | 用户之间友谊        | $G_t$          | 时序划分的子图 $t$  |
| $c$         | Check-in边     | $Z_U$          | 用户节点向量集合     |
| $N_v$       | 节点 $v$ 邻居节点集合 | $Z_A$          | POI节点向量集合    |
| $\vec{u}_i$ | 用户节点初始向量      | $W_{u,a}$      | 节点游走序列集合     |
| $w_t$       | 时间段 $t$ 的权重系数 | $Win$          | 游走序列窗口       |
| $adj$       | LBSN的邻接矩阵     | $\vec{z}$      | 节点 $v$ 向量表达  |
| $lv$        | 节点 $v$ 的游走序列  | $\mathbb{R}^d$ | 向量维度空间       |
| $\vec{a}_j$ | POI节点初始向量     | $x_v$          | 节点特征向量       |
| $\theta(v)$ | 用户节点的邻居集合     | $h(v)$         | 节点 $v$ 的表征向量 |

交网络层；Check-in边如图1中虚线所示，Check-in边构成了异构签到网络层。

**定义2.** 时序划分. 将LBSN中签到时间作为划分标准，依据集合  $T = \{t_1, t_2, \dots, t_T\}$  中不同的时间段将LBSN划分成  $T$  个子图，LBSN时序划分的子图表示为  $\{G_t\}_{t=1 \dots T}$ 。

以图5(b)为例，将LBSN依据时间段进行子图划分，从而将LBSN划分成  $T$  个不同的子图分别进行子图上的表示学习。

**定义3.** 邻居聚合路径. 给定  $G_L(V, E)$ ，在  $G_L$  中，以用户节点  $u$  为例，从用户节点  $u$  开始聚合邻居信息，邻居聚合路径为  $\langle u, a_1, u_2, a_3, \dots, u_i \rangle$ ，其中邻域内节点依据选择聚合邻居策略SAN进行选取。

**定义4.** 子图节点向量生成. 给定LBSN用  $G=(V, E)$  表示，其中  $V: \{U, A, T, L\}$ ， $E: \{e, c\}$  分别表示节点和边的集合，依据时序划分的时间段进行子图学习，捕捉用户节点之间的好友关系以及当前子图下用户节点与兴趣点之间的签到关系即用户流动性信息，将子图  $\{G_t\}_{t=1 \dots T}$  作为投影函数  $\{f_i\}_{i=1 \dots T}$  的输入，旨在获得节点投影生成维度为  $d$  的向量，且  $d \ll |V|$ ，子图的节点表示学习如公式(1)表示，其中  $Z_U$  和  $Z_A$  分别表示LBSN中得到的用户节点和兴趣点节点的向量表示。

$$f_i: \begin{cases} U \rightarrow Z_U, Z_U = \{\vec{z}_{U1}, \dots, \vec{z}_{Un}\}, \vec{z}_{Ui} \in \mathbb{R}^d \\ A \rightarrow Z_A, Z_A = \{\vec{z}_{A1}, \dots, \vec{z}_{An}\}, \vec{z}_{Ai} \in \mathbb{R}^d \end{cases} \quad (1)$$

**定义5.** 注意力机制. 借助GAT<sup>[23]</sup>计算权重

的方式，在共享权重参数  $W$  和节点特征向量  $\vec{h}_i$  的基础上，利用权重参数为  $\vec{a}$  的前馈神经网络，通过LeakyReLU函数进行非线性变换，计算节点  $v_i$  与邻居  $v_j$  间的权重系数如公式(2)，其中  $\parallel$  表示拼接操作：

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{a}^T [\vec{W}\vec{h}_i \parallel \vec{W}\vec{h}_j]))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(\vec{a}^T [\vec{W}\vec{h}_i \parallel \vec{W}\vec{h}_k]))} \quad (2)$$

### 3.2 异构网络表示学习

**定义6.** 异构网络. 给定  $G_{he}=(V, E, T)$ ，其中节点  $v$  和边  $e$  分别满足  $\varphi(v): V \rightarrow T_V$  和  $\varphi(e): E \rightarrow T_E$ ，其中  $T_V$  代表节点  $v$  的类型集合， $T_E$  代表关系  $e$  的种类集合，并且满足  $|T_V| + |T_E| > 2$ 。

**定义7.** 异构网络表示学习<sup>[26]</sup>. 给定异构网络  $G_{he}$ ，异构网络表示学习过程即将  $G_{he}$  作为输入，通过图表示学习学得向量表示，表示为： $f(G_{he}) \rightarrow X \in \mathbb{R}^{|V| \times d}$ ， $d \ll |V|$ ，其中学得  $X$  具有捕获异构网络中的结构信息和语义信息。

异构网络表示学习的过程就是依据边的类型进行划分，分别进行邻居聚合操作，假设现在异构网络  $G_{he}=(V, E, T)$ ，其中  $T_V = \{n_1, n_2\}$ ，且  $T_E = \{r_1, r_2\}$ 。由图2可知，节点  $u$  有两种类型的边，分别对应两种类型的节点，且与不同节点之间的关系包括  $r_1$  和  $r_2$ ，其中  $u \in n_2$ ， $e(n_2, n_2) \in r_1$ ， $e(n_1, n_2) \in r_2$ ，分别对节点  $u$  在  $r_1$  关系和关系  $r_2$  下的邻居进行聚合操作，利用神经网络训练得到对应的权重矩阵  $W_{r_1}$ ， $W_{r_2}$ ，从而得到由邻居节点特征向量加权变化而成的向量表示，并与节点自身的特征  $h_{r_1}(u)$ ， $h_{r_2}(u)$  向量共同学得节点  $u$  的向量表示。

### 3.3 问题定义

TGE-LBSN表示学习：借助注意力机制，依据特定下游任务，为子图  $\{G_t\}_{t=1 \dots T}$  的节点向量学得各时间段对应的自适应的权重值  $w_1, w_2, \dots, w_T$ ，从而得到LBSN的节点最终向量表示  $\vec{z} = w_1 * \vec{z}_1 + w_2 * \vec{z}_2 + \dots + w_T * \vec{z}_T$ ，其中  $\vec{z} \in \{Z_U, Z_A\}$ ，第  $i$  行代表第  $i$  个节点的  $d$  维向量表示。TGE-LBSN学习模型的优化目标是使得向量矩阵  $\vec{z}$  尽可能最大化的保留LBSN中的复杂结构信息和上下文语义信息。

以上是对一些基本的LBSN概念进行了形式化的定义表示，并作以详细说明，为下文TGE-LBAN框架介绍做了规范化的说明。



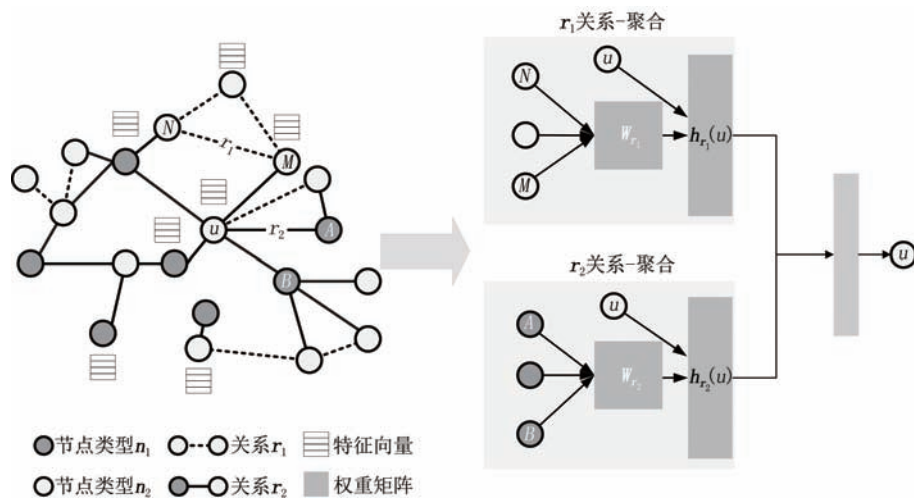


图2 异构网络表示学习过程

## 4 初始向量生成

本节主要对LBSN表示学习框架做了前期预处理,目的是通过阶段一的学习得到初始向量,并作为阶段二的输入.在接下来的小节中首先是对LBSN的Check-in数据进行有效低失真采样,并针对LBSN设计特定的游走策略,为初始向量生成做铺垫,以及阶段一初始向量生成阶段的详细介绍.

### 4.1 采样Check-in数据

Check-in数据包含了丰富的位置信息、时间信息、用户打卡地点的语义信息等,这使得Check-in数据成为LBSN区别于传统社交网络的关键信息,它不仅仅能有效刻画出用户的行为轨迹,还能辅助获得用户偏好设置.但LBSN中每个用户都包含了大量的Check-in数据,即LBSN实质上是异构超边网络,而聚合签到信息的过程无法保留全部签到信息,且用户的Check-in数据量差异很大,若采样固定数目的签到数据量,必然会损失多样性信息.本文提出一种有偏采样方式来抽取原始网络中的签到数据,对不同签到数量的用户节点给定不同的Check-in抽样数量,旨在高效地利用LBSN中丰富的Check-in数据,以挖掘LBSN中隐含的用户偏好,并将其保留在节点的向量表示中.

本文依据以下准则作为采样Check-in数据标准:(a)签到数据的抽取数目能表征节点之间的差异性;(b)用户最近的行为轨迹不仅能刻画此时用户当前的偏好甚至可以预测未来的行为轨迹,相反,时间越是久远的签到信息其参考价值可能相对较少.

首先,依据准则(a)对每个节点依据节点重要程

度确定采样超边的数量.用户节点将采样与其重要程度匹配的签到信息数量,节点越重要采样的Check-in边的个数越多.对于节点重要程度的衡量,综合考虑社交网络层次的度数以及Check-in数据数量、节点的PageRank作为评分依据,给出每个用户特定的采样比例,从而确定用户节点的Check-in数据的采样数量.其次,依据准则(b)确定选取Check-in数据的策略,采样策略满足时间越临近的超边,被采样的概率值越大.针对用户的所有Check-in数据按照时间线进行排序,日期临近的Check-in数据赋予较高的抽样概率,而日期相对久远的Check-in数据由于对用户现有偏好影响相对较小,则给定相对较低的抽样概率.图3(a)展示了依据时间顺序排列的同一用户签到信息,假设该用户有多条Check-in数据,从上到下是日期由远及近的顺序,本文设计的抽取策略分配给近期的Check-in数据所采样概率更大,即相较于活动点 $a_1, a_2$ ,对于用户 $u_1$ 而言,采样与活动点 $a_3$ 的Check-in数据概率更高.本文针对数据集签到时间跨度进行划分,将 $T$ 划分为 $n$ 个部分,假设第一条签到数据在第一部分则取值为1,如 $p_i = i / (1 + \dots + n)$ ,其中 $i$ 代表Check-in数据第 $i$ 个时间段,而后依据对应概率 $p_i$ 进行有偏抽取,从而确定当前用户对应的有偏采样的Check-in数据,得到筛选后的LBSN数据集.

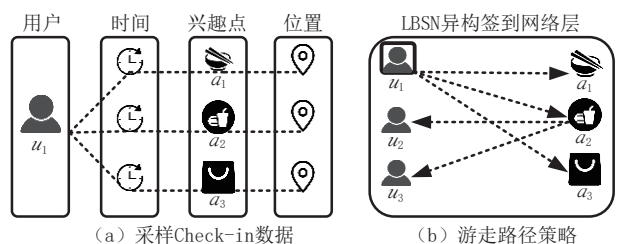


图3 有偏采样策略

## 4.2 LBSN上的游走策略

针对4.1节采样Check-in数据后的LBSN数据集,在生成节点初始向量之前,首先设定相应的游走路径策略,以确保向量表示可以学习到原始网络中特征信息以及好友关系,所以本文设定向量投影空间的节点相似性准则:在原始空间中关系相近的用户节点,在向量空间中的其向量表示也尽可能相似;在原始空间中属于同一条签到信息上的节点,也应该在投影的向量空间保持着签到关系,即在投影空间中向量之间距离保持相近.由于LBSN中存在两个层次的网络,社交关系层和签到异构层.对于用户节点来说,要获得初始的向量就需要包含用户之间的好友关系以及在异构层的签到信息.针对两个层次的网络,由图1可知,对于当前节点 $u$ 来说,存在社交网络层与用户之间的游走路径(图中红色实线路径),以及签到异构层的签到路径(蓝色路径).本文设计了游走策略:(a)用户与用户之间好友边的有偏选取,关系越紧密的用户,越有可能被选中,如图3(b)所示,对于当前用户节点 $u_1$ (黑色框)来说,二阶邻居 $u_2$ 和 $u_3$ 与 $u_1$ 的相似度存在差异,对于用户 $u_1$ 来说, $u_3$ 的轨迹与 $u_1$ 重合度更高,彼此更为相似;(b)用户与兴趣点之间频繁签到以及最近签到更有可能被选中,一阶邻居活动类别节点存在签到次数、签到时间等差异,签到越频繁时间越临近的活动节点更有可能被选取.

依据这两种游走策略,为每个节点依概率选取游走邻居,作为初始向量生成的依据.假设用户节点 $v$ 游走序列 $l_i = \{n_1, n_2, \dots, n_{k-1}\}$ ,若在社交网络层进行游走邻居选取公式(3)的第一个公式计算游走的转移概率 $p(n|l_i)$ ,即 $v$ 选择走向用户节点 $n$ 的概率.本文借助两个用户节点之间共同打卡活动点数目 $\pi_n$ 作为衡量用户节点亲密度的依据.若在签到异构网络层游走,则用公式(3)第二个公式选取游走路径上的活动节点,依据用户和活动点之间的签到次数 $c_{(v,n)}$ 以及最近签到时间 $t_{(v,n)}$ 来计算转移概率值 $p(n|l_i)$ ,其中 $p$ 和 $q$ 分别表示签到次数比重、签到时间比重,满足 $p+q=1$ .

$$p(n|l_i) \begin{cases} \frac{\pi_n}{\Delta_1} & \text{if } n \in U \\ \frac{p \cdot c_{(v,n)} + q \cdot t_{(v,n)}}{\Delta_2} & \text{if } n \in A \end{cases} \quad (3)$$

其中 $\Delta_1$ 和 $\Delta_2$ 用来做最大最小归一化处理,分别代表用户节点和活动节点对应计算内容的阈值大小,辅

助计算转移概率,分别对应公式(4)和公式(5),其中 $\theta(v)^k$ 代表节点 $v$ 的 $k$ 阶邻居节点集合.

$$\Delta_1 = \text{MAX}\{\pi_{v,n}\} - \text{MIN}\{\pi_{v,n}\}, n \in \theta(v)^k \quad (4)$$

$$\Delta_2 = \text{MAX}\{p \cdot c_{(v,n)} + q \cdot t_{(v,n)}\} - \text{MIN}\{p \cdot c_{(v,n)} + q \cdot t_{(v,n)}\}, n \in \theta(v)^k \quad (5)$$

## 4.3 IVGS算法

前两小节指定了LBSN上有偏采样策略和游走策略,本节将在采样后的数据上依据4.2节设定游走策略的两个准则,分别进行社交网络层的有偏游走 $S\_Walk$ 和签到异构层的有偏游走 $A\_Walk$ ,并利用两种类型的边,按照公式(3)计算不同的转移概率值完成有偏游走,分别得到用户节点游走序列 $W_u$ 及活动节点游走序列 $W_a$ ,为了生成初始的节点向量,本文利用SkipGram的思想,设定最小化损失函数来保证用户节点之间邻近度以及与活动类别节点之间的签到关系<sup>[18]</sup>,损失函数如下:

$$L_{loss} = -\log\left(\prod_{v \in V} \prod_{n \in \theta(v)} p(n|v)\right) \quad (6)$$

初始向量生成目的是为了确保原始图数据空间与投影的低维向量空间尽可能相似,并保证在节点特征信息融合的过程,保留原始网络中丰富的语义信息,将各层节点投影到同一个向量空间,从而得到初始特征向量的更新.初始化向量生成阶段过程参考算法1.算法1第1行利用4.1节有偏采样后的LBSN数据,第2行目的是获取节点初始向量,其中 $\vartheta: v \in V \rightarrow \mathbb{R}^{|V| \times d}$ ,第3~5行随机打乱节点顺序,调用对应游走函数 $S\_Walk$ 、 $A\_Walk$ 分别得到用户节点在社交网络层与异构网络层得到两种游走的序列,并将两种类型的游走序列保存在 $W_u$ .由图1可知对于兴趣点只需在异构网络层有偏游走,调用 $A\_Walk$ 得到活动节点对应游走序列(第9,10行),最后对于两种类型节点的游走序列,均调用SkipGram算法(算法2),以生成节点的初始向量 $\vec{u}, \vec{a}$ (第6,11行).

### 算法1. IVGS算法.

输入: LBSN 异构网络图  $G_L$ , 游走长度  $l$ , 窗口大小  $Win_u$

输出: 节点的初始向量  $\vec{u}, \vec{a}$

1.  $G \leftarrow \text{Sample}(G_L)$
2. Initialize sample  $\vartheta$  from  $V^{|V| \times d}$  // 取得初始向量
3. FOR each  $u$  in  $U$
4.  $U \leftarrow \text{Shuffle}(U)$
5.  $W_u \leftarrow S\_Walk(G_S, u, l) + A\_Walk(G_A, u, l)$
6.  $\vec{u} \leftarrow \text{SkipGram}(\vartheta, W_u, Win)$



7. END FOR
8. FOR each  $a$  in  $A$
9.  $A \leftarrow \text{Shuffle}(A)$
10.  $W_a \leftarrow A\_Walk(G_A, a, l)$
11.  $\tilde{a} \leftarrow \text{SkipGram}(\vartheta, W_a, \text{Win})$
12. END FOR

SkipGram算法(见算法2)主要利用自然语言处理的跳字模型为游走序列生成节点的向量表示,借助中心节点与路径节点之间出现的概率,引入最大似然估计设计损失函数(公式6),从而完成节点向量生成.由于损失函数包含节点数目的项的累加,对于LBSN的大规模节点来说,会导致每次计算的梯度开销过大.为了降低计算复杂度,本文采用层序softmax(Hierarchical softmax)做近似训练来加速SkipGram算法的训练过程,通过对条件概率进行因子分解以简化计算复杂度.层序softmax通过构建二叉树,将节点作为二叉树的叶子结点,即给定节点 $v$ 生成节点 $n$ 的条件概率计算就是将向量 $v$ 和二叉树根节点到 $n$ 路径上的非叶子节点向量进行一一求内积.具体来说,假设 $R(n)$ 代表从二叉树根节点到节点 $n$ 的路径上的节点数, $r(n, j)$ 代表该路径上的第 $j$ 个节点,并设该节点向量表示为 $u_{r(n, j)}$ ,那么利用Hierarchical softmax将SkipGram算法的条件概率可以近似表示为公式(7):

$$p(n|v) = \prod_{j=1}^{R(n)-1} \sigma(\llbracket r(n, j+1) = lc(r(n, j)) \rrbracket \cdot u_{r(n, j)}^\top v) \quad (7)$$

其中 $lc$ 表示判断是否为二叉树的左孩子,而 $\llbracket r(n, j+1) = lc(r(n, j)) \rrbracket$ 则表示为若路径上的当前节点是左孩子,则取值为1,反之为-1.除此之外,在训练过程中本文采用随机梯度下降(SGD)来优化调控参数,利用反向传播算法来估计导数,具体流程见算法2第4行.整个过程用以加速SkipGram算法处理游走序列以生成节点初始向量表示的过程. SkipGram算法如下:

**算法2.** SkipGram算法.

输入: 游走序列  $W_L$ , 窗口大小  $\text{Win}_u$

输出: 节点的初始向量  $\tilde{z}$

1. FOR each node  $v_n \in W_L$  in  $W_L$
2. FOR each  $k \in W_L[n - \text{Win}_u; n + \text{Win}_u]$
3. compute  $L_{\text{loss}}$  by (6)
4.  $\vartheta = \vartheta - \alpha * \frac{\partial L_{\text{loss}}}{\partial \vartheta}$
5. END FOR
6. END FOR

## 5 LBSN表示学习框架

在本节将着重介绍TGE-LBSN的图表示学习框架.首先5.1节概括性地介绍TGE-LBSN框架流程,详细阐述了框架构造的每个组成部分;5.2节提出阶段二选择聚合邻居的策略SAN算法;5.3节设计了TGE-LBSN模型的阶段二子图向量生成的SVG算法,得到子图下的节点向量;5.4节依据注意力机制调控各子图的权重系数,从而加权生成最终节点向量表示.

### 5.1 TGE-LBSN框架概述

本研究提出的TGE-LBSN框架主要包括两个阶段的LBSN图表示学习,图4将整个TGE-LBSN算法框架以流程图的形式展现出来,整体大致分为:数据集预处理模块、阶段一生成初始向量模块、阶段二最终向量生成模块.图中最上面部分是对数据集进行的预处理操作,利用设计的LBSN有偏采样策略,采样用户的Check-in数据,从而得到精简的LBSN数据集,而后进入阶段一(图4深灰色区域,第4节),通过设计对应游走策略生成了初始向量.对于4.3节得到的阶段一初始向量结合节点自身特征向量,作为阶段二(图4浅灰色区域)的输入数据.

为了更好地利用LBSN丰富的上下文语义信息,抓取LBSN网路动态演变的特性,阶段二首先借助签到时间进行LBSN数据的时序划分,利用数据之间的时序关系进行子图向量生成模型的训练.在各子图上分别进行邻居聚合的过程,并设计特定的聚合邻居选择策略SAN(见5.2节),以一种类似于元路径方式来规范聚合邻居路径,旨在学得异构时空网络中多时间段子图的节点向量表示,目的是在子图生成的节点向量(见5.3节)能最大化保留原始异构时空网络中的复杂结构信息及丰富的上下文语义信息,即用户之间的好友关系、用户与兴趣点之间的签到关系.

在子图向量汇总阶段,借助注意力机制学得自适应的权重系数,并利用设定的损失函数进行反向调控,将每个子图得到的嵌入表示进行加权聚合形成最终的向量表示(见5.4节),利用训练集的标记数据进行调控,并把面向训练任务的向量表示在测试集上进行精确度衡量.

### 5.2 邻居聚合选择策略

面对大规模的复杂图数据,图的节点数存在指数级增长的问题,如若图结构中的平均节点度数用

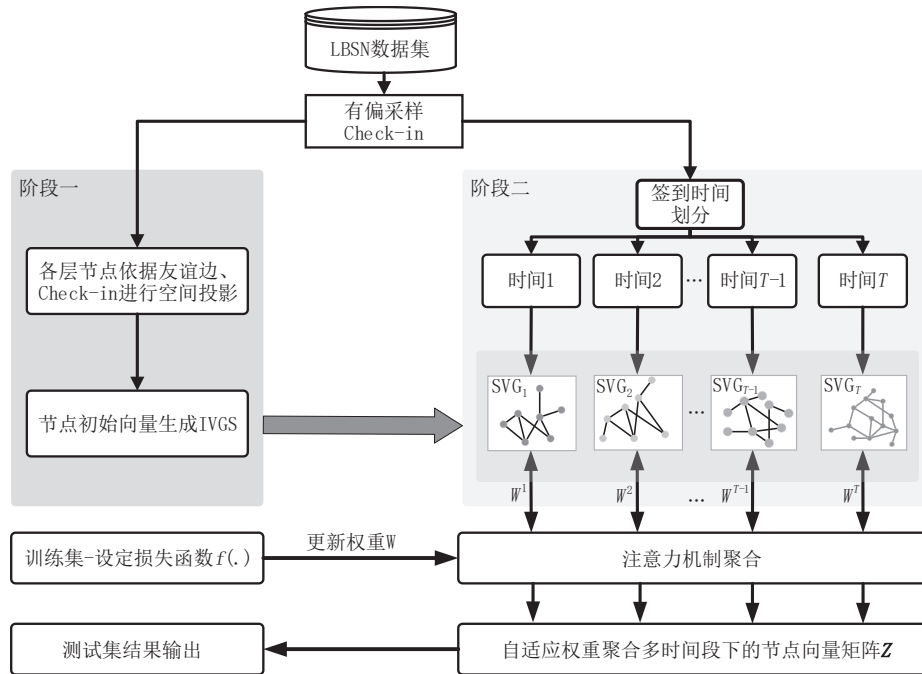


图4 TGE-LBSN框架总图

$\bar{d}$ 来表示,当执行 $K$ 阶聚合操作,那么 $K$ 阶子图将平均包含 $1 + \bar{d} + \bar{d}^2 + \dots + \bar{d}^k$ 个节点.当 $d$ 和 $k$ 比较大的时候,所需要进行聚合邻居的数目非常之大,计算复杂度极高.除此之外,大部分的图数据节点的度往往呈现幂律分布,在超级节点聚合邻居的过程中会面临着邻居暴增的可能,即聚合超级节点的 $K$ 阶邻居时<sup>[22]</sup>会出现邻居数量巨大的现象.所以对于LBSN这类大型异构时空网络,同样在子图聚合邻居的过程中出现 $K$ 阶邻居暴增的现象,这就需要对所聚合的邻居做出进一步的选择,在同一阶邻居聚合的过程中,确定聚合邻居的数量以及设计合理的邻居选择方式显得至关重要.不同阶层的邻居对当前节点的向量生成存在差异性.图3(b)展示了节点在聚合过程中,同阶邻居与当前节点的亲密度存在差异性,所以须对 $K$ 阶邻居依据差异性进行选择聚合.

本文为了减少聚合邻居的数目,避免出现聚合邻居数量暴增的现象,且要尽可能保留原始异构网络中节点之间的邻近关系,选择能充分表征当前聚合节点的部分邻居完成聚合操作.由于图表示学习更倾向于保留原始网络中节点之间的邻近关系,本文将对邻居节点进行选择性聚合,选定满足要聚合数目的相似邻居节点作为聚合邻居完成当前节点的聚合操作.图5(a)展示了LBSN对于用户节点 $u$ 以及兴趣点 $c$ 的聚合路径(分别用黑色箭头和灰色箭头表示),在聚合邻居过程中采用特定邻居选择策略

来决定聚合邻居.在聚合过程对于邻居选择会倾向于选择与当前聚合节点更相似的邻居节点,以便能更好的表征当前节点的特征,并学得可以表示节点之间相似关系的节点向量.所以本文设计了邻居选择策略SAN(算法3),对聚合的邻居进行有偏抽取,利用选择后的邻居节点完成聚合操作,而非采用随机采样的方式进行邻居聚合.

算法3提供了当前节点聚合邻居的选择算法,首先对于当前节点的所有邻居,计算其与邻居之间的签到次数,也就两个节点之间边的个数,以及最近的签到时间,用签到时间和签到次数作为两个基本的衡量标准(第2,3行),而后对邻居节点的类别进行划分,若是当前节点是用户节点,聚合邻居考虑的因素除了签到时间、签到次数之外,还须考虑签到点的人流量(第5行),并利用 $w_{vi} = \alpha_u * time + \beta_u * sum + \gamma_u * d$ 计算出当前节点的所有邻居对应的 $w_{vi}$ ,其中 $\alpha_u, \beta_u, \gamma_u$ 分别代表这三个指标的各自所占比重,且满足 $\alpha_u + \beta_u + \gamma_u = 1$ (算法第6行);若是活动类别节点,还需要考虑是否存在前一个节点,确定前一个节点与即将聚合的节点之间的初始向量的相似度 $\delta$ ,作为另一个衡量依据(第8-17行),同样 $\alpha_u, \beta_u, \gamma_u, \delta$ 是对应评价指标的权重系数,满足 $\alpha_u + \beta_u + \gamma_u + \delta = 1$ .随后汇总每一个邻居节点计算出来的 $w_{vi}$ ,第19行做归一化处理得到 $\omega$ ,作为选择邻居的依据,利用计算好的概率值进行有偏抽取

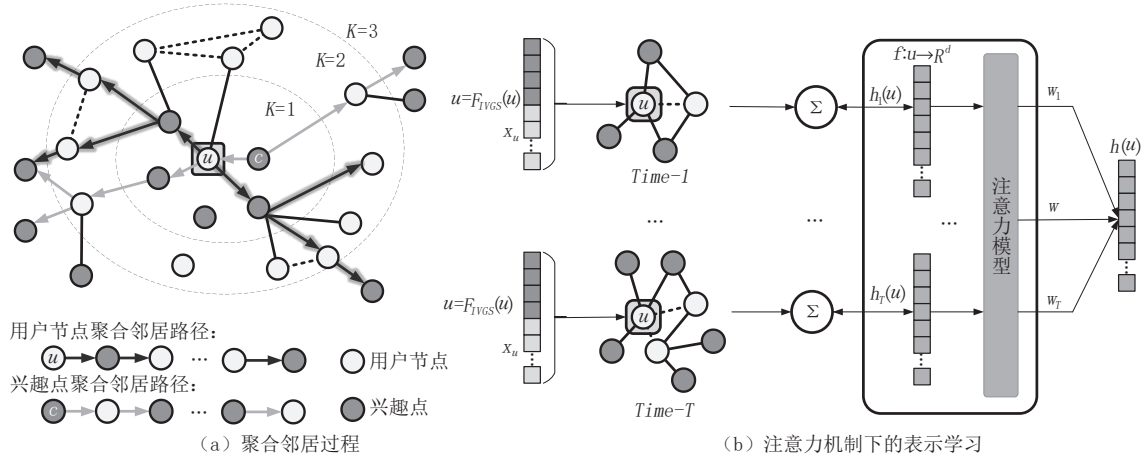


图5 LBSN节点向量生成

(第20行),最终得到节点 $v$ 选中的邻居节点集合并将结果返回(第21行).

### 算法3. SAN算法.

输入: LBSN异构网络图 $G_A$ 邻接矩阵 $adj$ ,当前节点 $v$ ,

Check-in数据信息 $C$

输出: 节点选择聚合的邻居 $\theta(v)$

1. FOR each  $i$  in  $adj(v)$ :
2.  $time = \max Time(C_{time}([i, v]))$
3.  $sum = \text{Number of occurrences}(i, v) \text{ in } C$
4. IF  $v \in U$
5.  $d = \text{Degree}_{G_A}(i)$
6.  $w_{vi} = \alpha_u * time + \beta_u * sum + \gamma_u * d$
7. END IF
8. ELSE
9.  $d = \text{Degree}_{G_S}(i)$
10. IF  $Pre(v)$  is not none:
11.  $si = \text{Similarity}(Pre(v), i)$
12. END IF
13. ELSE
14.  $si = 0$
15. END ELSE
16.  $w_{vi} = \alpha_a * time + \beta_a * sum + \gamma_a * d + \delta * si$
17. END ELSE
18. END FOR
19.  $\omega = \text{Normalization}(w_{vi}, i \in adj(v))$
20.  $\theta(v) = \text{Biased aggregation}(\omega) // \text{依据 } \omega \text{ 抽取}$
21. RETURN  $\theta(v)$

### 5.3 子图向量生成SVG算法

为了保留LBSN中的签到关系和复杂结构信息,针对LBSN中多类型的节点形成的异构时空网络,阶段二的主要是将异构网络中的复杂的结构信息融合到阶段一的初始向量中.同时为了充分利用

LBSN数据的时序特性,首先将原始复杂的异构时空网络,依据签到时间进行多时间段子图划分,整体流程可以用图5(b)来概括,图中以节点 $u$ 为例,节点 $u$ 出现在多个时间段下的子图中, $u$ 节点的向量生成首先利用阶段一的向量及自身特征向量拼接成阶段二的初始向量,并作为各个子图模型的输入,并针对每个时间段对应的子图进行异构网络上的用户节点及兴趣点的邻居聚合操作SAN算法,从而得到多时间段下各子图聚合复杂结构信息和签到语义信息的节点向量 $\{h_1(u), h_2(u), \dots, h_T(u)\}$ .

SVG算法(算法4)展示了在获取到阶段一节点初始向量的基础上,聚合子图内邻域信息并学得节点向量表示的过程.第1~3行将阶段一的初始节点向量与特征向量进行拼接,作为子图聚合邻居的模型输入. SVG算法搭建K层的网络层,即考虑节点K阶邻居的聚合(第4行).对于子图中每个节点无论是用户节点还是兴趣点(第5行),对这两种类型的节点开始邻居聚合的过程,调用5.2节的邻居选择策略SAN来确定聚合邻居(第6行).在异构网络邻居聚合过程中,异类随机游走倾向于走向高度可见的节点类型以及集中的节点<sup>[36]</sup>.由图5(a)可知,针对用户节点 $u$ ,对邻域内的兴趣点 $a$ 进行SAN抽样,类似地,兴趣点选择打卡过的用户作为下层聚合邻居,同时当前节点是用户节点 $u$ 与当前节点是兴趣点 $a$ 时聚合路径是不同的,分别对应图上黑色路径和灰色路径,且SAN决定邻域内邻居的选择策略,而 $\alpha_{uj}^k$ 和 $\alpha_{aj}^k$ 决定了已选邻居在聚合邻居过程中所占比重(第7行),即在邻域内选定的邻居匹配自适应的权重,并利用公式(8)计算聚合选中的邻居得到的表征向量,分别得到邻居的聚合层 $h_{SAN(u)}^k$ 和



$h_{SAN(a)}^k$  (第9行), 其中AGG聚合的形式可以为平均或加和聚合算子, 如公式(9), 也可替换为池化算子, 如最大池化操作, 如公式(10).

$$h_{SAN(v)}^k \leftarrow AGG_{ku} \left( \left\{ \alpha_{vj}^k w^k h_j^{k-1}, \forall j \in SAN(v) \right\} \right) \quad (8)$$

$$Agg^{sum} = \sigma \left( SUM \left\{ Wh_j + b, \forall v_j \in N(V_i) \right\} \right) \quad (9)$$

$$Agg^{pool} = MAX \left\{ \sigma(Wh_j + b), \forall v_j \in N(V_i) \right\} \quad (10)$$

由于聚合邻域节点的特征信息的同时能捕捉异构网络层的结构信息, 而节点自身的信息包括自身特征信息以及阶段一得到用户间的友谊信息, 将聚合的邻居信息与节点自身的特征信息  $h_v^{k-1}$  进行拼接作为本层的节点向量表示(第10行), 再利用非线性激活函数得到当前子图下节点的向量表示  $h_v^k$  (第12行), 最后得到  $K$  层节点向量表示集合  $\{h_v^k, k=1, \dots, K\}$  作为SVG算法输出(第14行).

#### 算法4. SVG算法.

输入: 时间段划分的子图  $G_t, t \in \{1, 2, \dots, T\}$ , 用户节点特征向量  $\{x_u, \forall u \in U\}$ , 兴趣点特征向量  $\{x_a, \forall a \in A\}$ , 神经网络深度  $K$ , 非线性激活函数  $\sigma$ , 邻居聚合方式  $AGG_k, \forall k \in \{1, \dots, K\}$

输出: 子图下节点的  $\{h_v^k, k=1, \dots, K\}$

1.  $\vec{u}, \vec{a} \leftarrow IVGS(G_t, l, Win_u)$
2.  $h_u^0 \leftarrow CONCAT(\vec{u}, x_u), \forall u \in U$
3.  $h_a^0 \leftarrow CONCAT(\vec{a}, x_a), \forall a \in A$
4. FOR  $k=1 \dots K$
5. FOR  $v \in V$
6. FOR  $j \in SAN(v)$
7. Compute  $\alpha_{vj}^k$  use (2)
8. END FOR
9. Compute  $h_{SAN(v)}^k$  by (7)
10.  $h_v^k \leftarrow \sigma(W_{kv} \cdot CONCAT(h_v^{k-1}, h_{SAN(v)}^k))$
11. END FOR
12.  $h_v^k \leftarrow \frac{h_v^k}{\|h_v^k\|_2}, \forall v \in V$
13. END FOR
14. RETURN  $\{h_v^k, k=1, \dots, K\}$

#### 5.4 基于注意力机制的TGE-LBSN算法

利用SVG算法得到子图的节点向量表示  $h_v^k$  时, 需要对各子图的节点向量表示, 采用注意力机制进行融合汇总, 为不同子图学得自适应的权重来合并多个子图模型下的节点向量得到最终的节点向量表示.

算法5展示了TGE-LBSN算法借助注意力机制融合多时间段下的节点向量的过程. 算法首先对于划分的  $T$  个子图部分(第1行), 调用SVG(算法

4)得到当前子图  $G_t$  对应的节点向量表示(第2行). 第3~4行分别对用户节点和兴趣点进行  $K$  层节点向量生成, 并加入非线性激活函数得到  $h_u^{K,t}, h_a^{K,t}$ . 在多个时间段下得到不同的向量表示, 由于注意力机制本质上就是为表征中比较有价值的部分分配较多的权重, 本文在多时间段的模型后引入注意力机制, 用以学得  $w_1, \dots, w_T$  权重系数. 由图5(b)可知, 借助注意力机制得到各个时间段所对应权重来融合多个时间段下的节点向量表示, 依据下游任务调控注意力机制, 分配给不同时间段的不同权重占比, 按权重加权求和的方式得到统一的最终节点  $u$  的向量表示(第5~6行), 从而得到最终的向量表示  $\vec{z}$  (第8行).

#### 算法5. TGE-LBSN算法.

输入: 时序子图  $G_t$  下的节点嵌入表示  $\{h_v^k, k=1, \dots, K\}$ , 非线性激活函数  $\sigma$

输出: 节点的最终向量  $\vec{z}$

1. FOR  $t=1 \dots T$
2. Compute  $SVG(G_t)$
3.  $h_u^{K,t} = \sigma \left( \frac{1}{K} \sum_{k=1}^K \sum_{j \in SAN(u)} \alpha_{uj}^k \cdot W_{ku} \cdot h_u^k \right)$
4.  $h_a^{K,t} = \sigma \left( \frac{1}{K} \sum_{k=1}^K \sum_{j \in SAN(a)} \alpha_{aj}^k \cdot W_{ka} \cdot h_a^k \right)$
5.  $H_u = w_1 * h_u^{K,1} + \dots + w_T * h_u^{K,T}$
6.  $H_a = w_1 * h_a^{K,1} + \dots + w_T * h_a^{K,T}$
7. END FOR
8. RETURN  $\vec{z} \leftarrow (H_u, H_a) \forall u \in U, \forall a \in A$

## 6 实验和评估

为了验证TGE-LBSN算法有效性, 本文选取了四种经典LBSN的真实数据集和四个时序社交网络数据集作为算法的数据集, 并从社交领域的好友预测和兴趣点预测两个任务上, 分别对LBSN异构网络表示学习的算法进行有效性验证, 并对预测任务的结果进行详细地分析说明, 最后通过消融测试证明了本研究所提创新点对于LBSN的信息捕获至关重要.

### 6.1 数据集及实验配置

本节主要介绍了实验相关数据集以及一些基准对比算法、评价指标、实验环境、参数设置等等一些基础实验配置.

#### 6.1.1 实验数据集

本研究数据集采用了LBSN真实的数据集以及时序社交数据集两大类, 第一部分是经典的LBSN数据集, 主要包括: Yelp, Gowalla, Foursquare, Brightkite 四种经典的LBSN数据集. 这些数据集主要包括用

户之间的社交网络层以及用户与兴趣点之间的签到网络,且用户与兴趣点之间的异构边具有时间属性,除此之外用户节点和兴趣点都具有位置等属性信息.第二部分是时序社交数据集,本文选取IAEnronEmployees (IAEE)、Collegemsg (ColM)、IAContact (IA)、IAContactsHypertext (IACH)四个作为时序社交网络数据集的代表.这些数据集主要包括用户之间的友谊信息,以及用户之间通信时间信息,与LBSN数据集相比,时序社交网络不包括用户的Check-in信息,仅包含社交网络层.

(1) LBSN数据集的详细信息见表2,其中Yelp数据集是典型的LBSN数据集,主要包含30 887个用户节点及860 888条签到数据信息;Gowalla是由18 737个节点和1 278 274个Check-in边组成;Foursquare该数据集包含2 153 469个用户,1 021 966个签到组成;Brightkite是LBSN服务提供商,由58 228个节点和4 491 143个Check-in边组成.

表2 LBSN数据集

| LBSN        | Yelp | Gowalla | Foursquare | Brightkite |
|-------------|------|---------|------------|------------|
| 用户节点        | 30k  | 19k     | 2153k      | 58k        |
| 兴趣点         | 19k  | 32k     | 1143k      | 57k        |
| Check_ins   | 861k | 1278k   | 1022k      | 4491k      |
| Friendships | 266k | 87k     | 27 098k    | 4747k      |
| 时间跨度(天)     | 370  | 375     | 372        | 791        |

(2) 时序社交数据集的详细信息见表3,其中IA-EnronEmployees (IAEE)共有50 571条数据信息;Collegemsg (ColM)有1899个节点和20 296个边;IAContact (IA)共有28 244条边;IAContactsHypertext (IACH)共有20 818条数据.

表3 时序社交网络数据集

| 时序社交网络      | IAEE | ColM | IA  | IACH |
|-------------|------|------|-----|------|
| 用户节点        | 0.1k | 2k   | 2k  | 0.1k |
| Friendships | 3k   | 20k  | 30k | 21k  |
| 时间跨度(天)     | 197  | 193  | 398 | 265  |

### 6.1.2 实验对比算法

除了在多个数据集验证所提出的算法之外,本文还与异构网络表示学习的一些算法、经典的表示学习等算法进行对比分析.由于本文要使用的数据集是LBSN,属于异构时空网络,由于经典算法本身并未处理位置时间等数据,以及异构超边等信息,实验过程对经典的表示学习算法对输入的数据集作了

相应的改动,以下是三种方法基准方法的介绍:

(1) 随机游走算法:DeepWalk<sup>[16]</sup>,LINE<sup>[17]</sup>和node2vec<sup>[18]</sup>等等,这类算法仅针对同构网络,且已经在文献[7]做对比实验,且由于这些算法无法使用节点的属性信息,即TGE-LBSN算法阶段一产生的初始向量无法作为输入,所以本实验在这一部分并未相关做对比实验.

(2) 深度学习图神经网络算法:以GCN<sup>[20]</sup>,GraphSAGE<sup>[22]</sup>、GAT<sup>[23]</sup>为代表,由于这些方法都是在同构网络上的图表示学习,本研究在实验过程中为了方便进行对比实验,在用户签到信息的异构网络层做兴趣点预测实验时做了相应同构网络的转化,并与GCN、GraphSAGE、GAT进行对比实验.

(3) LBSN异构嵌入技术:LBSN2vec++<sup>[7]</sup>,该算法主要是DeepWalk的思想应用在LBSN网络上,但是并未考虑节点自身初始的特征向量.本文在做相关对比实验时,通过与LBSN2vec++算法对比,来间接对比了其他随机游走算法.

### 6.1.3 评价指标

对于社交领域的各项预测任务如好友预测、兴趣点推荐等,本文主要采用ROC(Receiver Operating Characteristic)曲线作为评价依据,利用真正例率TP、假正例率FPR,其中横轴FPR越大,预测正类中实际负类越多,纵轴TPR越大,代表预测正类中实际正类越多,利用ROC曲线图进行结果展示,其中曲线下面积越大,表示AUC(Area Under roc Curve)值越大,预测精确度越高.

除此之外,由于好友预测和兴趣点预测,本质上是链路预测,即判断两个节点之间是否存在一条边,可以视为0-1分类问题,所以本文还采用了准确度(Accuracy)、精确率(Precision)、召回率(Recall)、F1值进行评估( $F_1$ -value),各个评价指标的具体计算公式分别对应公式(11~14),其中,TP代表将正类预测为正类,FP是把负类预测为正类,TN代表负类被判定为负类.FN代表正类被判定为负类.如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (14)$$

#### 6.1.4 实验环境及参数设置

**实验环境：**64位操作系统：Windows10, CPU: i7-10700@2.90 GHz, 内存为16 GB, 硬盘500 GB, 编程环境为Python 3.8, Pytorch1.6.0.

**参数设置：**对于TGE-LBSN实验以及对比算法给出相应实验参数设置如下：

(1) **TGE-LBSN：**对于数据集划分比例  $data\_split=[0.3, 0.2, 0.25, 0.25]$ . 由于TGE-LBSN主要包括两个阶段的参数设置, 第一阶段需要设置的参数如下: 在通过SkipGram生成节点向量时, 窗口大小设置为5; 节点的游走步长  $l$  设置为20; 节点  $u$  生成的向量  $\vec{z}_u$  的维度大小  $d=100$ , 游走策略中  $p$  和  $q$  经过大量实验测试证实了当  $p=0.7, q=0.3$  时得到的初始向量对阶段二的性能提升更有帮助. 阶段二中参数设置为: 学习率  $r=5e-4$ , 分类任务中的损失函数为交叉熵损失函数; 丢弃率  $weight\_decay=5e-3$ ; 在SAN算法中针对邻居的评价指标系数设置如下, 对于用户节点考虑超参  $\alpha_u, \beta_u, \gamma_u$ , 由于  $\alpha_u + \beta_u + \gamma_u = 1$  且  $\alpha_u, \beta_u, \gamma_u > 0$ , 此处实验设置为  $\alpha_u \in (0, 0.5)$ ,  $\beta_u \in (0, 0.5)$ ,  $\gamma_u = 1 - \alpha_u - \beta_u$ , 经过在真实数据集上进行测试参数敏感度, 在实验过程中, 结果表明在大多数情况下  $\alpha_u \in (0.3, 0.5)$ ,  $\beta_u \in (0.3, 0.5)$ ,  $\gamma_u = 1 - \alpha_u - \beta_u$  时多数情况下实验性能会取得最优, 所以实验在这部分参数设置为:  $\alpha_u = 0.4, \beta_u = 0.4, \gamma_u = 0.2$ ; 对于活动节点的超参设置主要包括  $\alpha_a, \beta_a, \gamma_a, \delta$  同理这四者之间满足  $\alpha_a + \beta_a + \gamma_a + \delta = 1$  且  $\alpha_a, \beta_a, \gamma_a, \delta > 0$ . 此处实验设置为:  $\alpha_a \in (0, 0.3)$ ,  $\beta_a \in (0, 0.3)$ ,  $\gamma_a \in (0, 0.3)$ ,  $\delta = 1 - \alpha_a - \beta_a - \gamma_a$ , 同理, 在实验中进行大量测试, 结果表明在  $\alpha_a \in (0.2, 0.3)$ ,  $\beta_a \in (0.2, 0.3)$ ,  $\gamma_a \in (0.2, 0.3)$ ,  $\delta = 1 - \alpha_a - \beta_a - \gamma_a$  时实验结果在多数情况下表现优良性能, 所以本文选取  $\alpha_a, \beta_a, \gamma_a, \delta = 0.25$  作为SAN算法中活动节点计算过程中超参的设置.

(2) **对比算法：**对于GCN<sup>[20]</sup>, 实验中层数设置为2层, 隐藏层个数为64, 学习率  $r=5e-4$ , 分类任务中的损失函数为交叉熵损失函数; 丢弃率  $weight\_decay=5e-3$ ; GraphSAGE<sup>[22]</sup> 实验中层数设置为2层, 隐藏层个数为128, 学习率  $r=5e-3$ , 分类任务中的损失函数为交叉熵损失函数; 丢弃率  $weight\_decay=5e-3$ ; GAT<sup>[23]</sup> 实验中层数设置为2层, 隐藏层个数为32, 学习率  $r=5e-3$ , 分类任务中的损失函数为交叉熵损失函数; 丢弃率  $weight\_decay=5e-3$ ; 迭代次数为3; LBSN2

vec++<sup>[7]</sup> 实验中节点嵌入维度为128, 随机游走参数游走长度  $l=80$ , 每个节点游走序列个数为10, 滑动窗口大小为10, 负采样  $\gamma=10$ .

#### 6.2 好友预测

好友预测是对LBSN的社交网络层的用户节点之间的是否存在好友关系进行预测. 在数据集的处理上, 利用  $data\_split$  对LBSN数据集依据时间进行划分为  $n$  个数据集, 如  $T_1, T_2, \dots, T_n$  数据集, 满足时间关系:  $T_1 < T_2 < \dots < T_n$ , 根据  $T_1, T_2, \dots, T_{i-1}$  作为训练集进行表示学习模型训练, 得到用户节点的嵌入表示, 再利用节点向量表示在  $T_i, \dots, T_n$  数据集中进行好友关系预测. 本文采用半监督模式对用户之间在社交网络层的好友关系进行预测. 通过文献[7]可知, 在好友预测时, 用户的社交信息更加重要, 在自身特征信息与初始特征向量拼接成新的初始向量应该在异构网络聚合过程中占据重要地位, 即友谊信息、自身特征信息相对于结构信息来说对用户之间的好友预测更有帮助. 这区别于兴趣点预测任务, 不是更注重聚合异构网络的结构信息, 所以阶段一得到的初始向量对于好友预测来说相对更重要, 详见6.4节消融测试. 由于时序社交网络缺乏位置信息, 而LBSN2vec++向量生成需要位置信息, 所以实验在这一部分并未进行对比实验.

本文为了验证TGE-LBSN算法能有效地捕捉LBSN网络上的友谊信息, 同时在真实的LBSN数据集和时序社交网络数据集上做了对比实验, 实验结果分别展示在图6和图7, 其中图6展示了TGE-LBSN算法与其他基准算法在四个真实的LBSN数据集上好友预测精确度对比. 图中直观地展示了在好友预测任务上, TGE-LBSN算法预测性能在4个数据集上均在一定程度超过了基准算法, 其中提升最多是TGE-LBSN在FourSquare数据集上最高可提升42.1%, 这也说明了TGE-LBSN在对LBSN聚合信息方便相比于GraphSAGE、GAT、GCN等学得的节点表示更能捕捉LBSN的友谊信息. 可见GraphSAGE、GAT、GCN在学习LBSN中, 由于损失了LBSN中的关键的上下文语义信息, 并不能充分挖掘LBSN中的友谊信息. 除此之外, 图7展示了本文在时序社交网络上进行好友预测的结果, 经过对比可知, TGE-LBSN在时序社交网络上的性能提升并不如在LBSN数据集上不明显, 最高提升是在ColM数据集上提升4.85%, 由此可见, 针对LBSN设计的图表示学习算法TGE-LBSN更适用于在



LBSN数据集上学习节点向量表示, TGE-LBSN所提策略对于LBSN更加敏感的捕捉到LBSN上复杂的结构信息以及上下文语义信息.

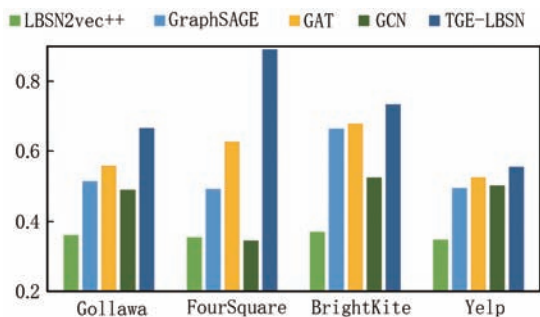


图6 LBSN好友预测结果图

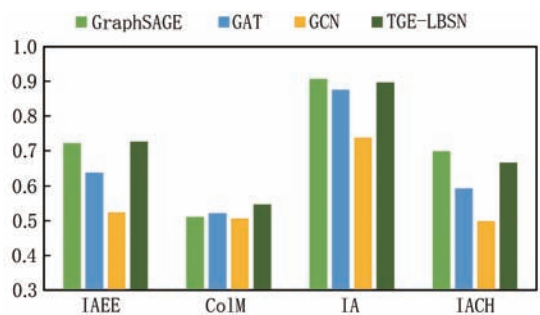


图7 时序社交网络的好友预测结果

为了更好地展示实验结果, 本文还绘制了好友预测在四个数据集上的ROC曲线以及精确率召回率PR曲线 (Precision-Recall Curve), 图8展示了TGE-LBSN算法在四个真实的LBSN数据集上好友预测的PR曲线, 图9展示了TGE-LBSN算法进行好友预测的ROC曲线展示. 通过四个数据集上的对比, 可以发现算法在FourSquare数据集上表现得尤其出色, 比在Yelp数据集上预测准确度高出60.56%, 可见TGE-LBSN在好友预测上更能捕获FourSquare数据集上的节点特征及丰富的语义信

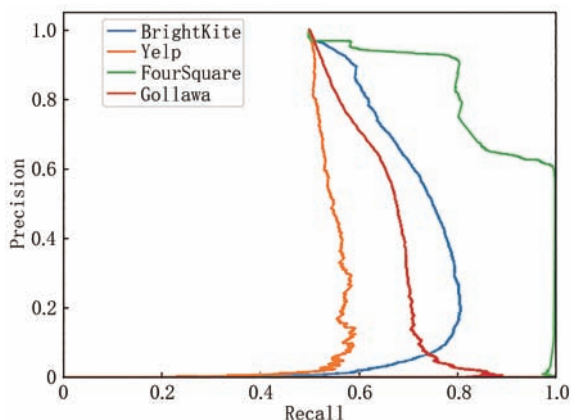


图8 好友预测PR曲线图

息. 除了AUC值, 本文还利用其它指标(准确率、精确率、召回率、F1值)与对比算法中的GAT、GCN在四个数据集上进行了对比. 由表4可知, TGE-LBSN在各种评价指标下都表现出不错的预测效果.

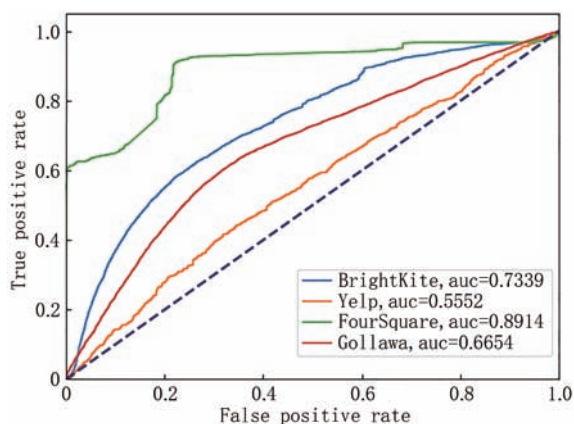


图9 好友预测ROC曲线图

### 6.3 兴趣点预测

兴趣点预测是在未来的一段时间内预测用户节点和兴趣点之间是否存在一条签到边. 在社交领域的推荐任务中, 兴趣点预测通过挖掘用户偏好, 对用户偏好做出行为分析, 预测用户即将喜欢或者即将签到打卡的兴趣点, 这对完成社交领域高质量推荐服务至关重要. 本文依据签到时间对用户节点和兴趣点之间的签到边进行训练得到的学习模型得到用户节点和兴趣点的向量, 用来预测未来一段时间内用户是否会访问给定的兴趣点. 实验过程是将数据集依据时间进行划分, 利用前一段时间的签到数据进行训练, 再利用下一时间段的数据作为测试集, 进行兴趣点预测.

为了更好地展示实验结果, 本文在多个真实LBSN数据集上, 利用AUC值作为评价指标, 与各个基准算法进行对比, 由图10可见TGE-LBSN算法在兴趣点预测方面, 在各个数据集上都取得了优秀的预测精确度, 且通过对实验结果进行分析, 在Yelp数据集上预测准确度高出7%, 以及在四个数据集上均对于基准算法有不同程度的提升, 这些结果都表明了本文设计的图表示学习过程对于节点的向量表示保留了较多有效的信息. 对于图10中TGE-LBSN相对于GAT、GraphSAGE提升并不明显这一现象, 在实际训练过程中TGE-LBSN训练速度优有所提升, 至于兴趣点预测精确度方面, TGE-LBSN对于活动节点的学习仅有异构签到层

| 表 4 好友预测的各指标对比 |          |          |           |        |          |               |
|----------------|----------|----------|-----------|--------|----------|---------------|
| 数据集            | 模型       | Accuracy | Precision | Recall | F1-value | AUC           |
| Yelp           | GAT      | 0.4995   | 0.2499    | 0.4995 | 0.3331   | 0.527         |
|                | GCN      | 0.5      | 0.25      | 0.5    | 0.3333   | 0.5028        |
|                | TGE-LBSN | 0.4981   | 0.6252    | 0.5008 | 0.3357   | <b>0.5552</b> |
| Gollawa        | GAT      | 0.4996   | 0.2499    | 0.4995 | 0.3331   | 0.5583        |
|                | GCN      | 0.4993   | 0.2498    | 0.4993 | 0.333    | 0.4905        |
|                | TGE-LBSN | 0.5159   | 0.6053    | 0.5159 | 0.3855   | <b>0.6654</b> |
| FourSquare     | GAT      | 0.4927   | 0.2482    | 0.4927 | 0.3330   | 0.627         |
|                | GCN      | 0.4997   | 0.4166    | 0.4997 | 0.3338   | 0.3768        |
|                | TGE-LBSN | 0.8435   | 0.8501    | 0.8435 | 0.8427   | <b>0.8914</b> |
| BrightKite     | GAT      | 0.4998   | 0.2499    | 0.4998 | 0.3332   | 0.6791        |
|                | GCN      | 0.5      | 0.25      | 0.5    | 0.3333   | 0.5259        |
|                | TGE-LBSN | 0.6438   | 0.6921    | 0.6438 | 0.62     | <b>0.7339</b> |

的信息支持,相较于用户节点信息获取比较单一,所以在这部分的实验并未表现大幅度的提升.

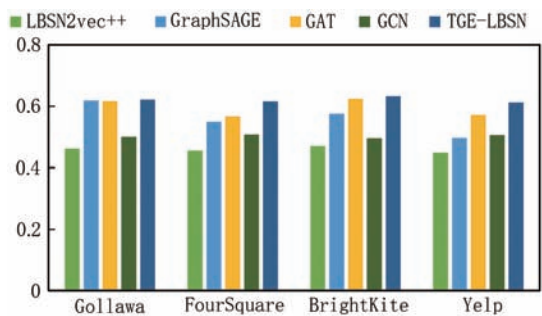


图 10 兴趣点预测对比结果图

本文进一步与 GAT、GCN 在准确率、精确率、召回率、F1 值上进行多指标对比,对应结果都展示在表 5,通过观察可以发现在多个指标评价标准上,TGE-LBSN 依旧能保持很好的效果.在 GAT、GCN 上效果相对比较低,因为这两类算法对于用户节点和兴趣点并没有进行分别对待,对于这两种节点的向量生成采用同样的处理方式,这一定程度上损失了 LBSN 特有的时空、语义信息.同时实验结果再一次验证了 TGE-LBSN 算法能捕捉到 POI 节点自身特征信息,以及异构网络层的复杂结构信息.

| 表 5 兴趣点预测的各指标对比 |          |          |           |        |          |               |
|-----------------|----------|----------|-----------|--------|----------|---------------|
| 数据集             | 模型       | Accuracy | Precision | Recall | F1-value | AUC           |
| Yelp            | GAT      | 0.5      | 0.25      | 0.5    | 0.3333   | 0.5718        |
|                 | GCN      | 0.5      | 0.5       | 0.5    | 0.3333   | 0.5019        |
|                 | TGE-LBSN | 0.5001   | 0.259     | 0.5    | 0.3332   | <b>0.6118</b> |
| Gollawa         | GAT      | 0.5000   | 0.2498    | 0.5000 | 0.333    | 0.6168        |
|                 | GCN      | 0.5      | 0.25      | 0.5    | 0.3333   | 0.5039        |
|                 | TGE-LBSN | 0.5      | 0.25      | 0.5    | 0.3333   | <b>0.6214</b> |
| FourSquare      | GAT      | 0.5034   | 0.5405    | 0.5034 | 0.3553   | 0.5667        |
|                 | GCN      | 0.5      | 0.5       | 0.5    | 0.3345   | 0.5077        |
|                 | TGE-LBSN | 0.5441   | 0.6495    | 0.5342 | 0.4228   | <b>0.6156</b> |
| BrightKite      | GAT      | 0.496    | 0.249     | 0.486  | 0.3315   | 0.6236        |
|                 | GCN      | 0.4997   | 0.2499    | 0.5    | 0.3332   | 0.5024        |
|                 | TGE-LBSN | 0.554    | 0.6488    | 0.554  | 0.4695   | <b>0.6323</b> |

本文绘制了兴趣点预测在四个数据集上的 ROC 曲线以及精确率召回率 PR 曲线,图 11 展示了 TGE-LBSN 算法在四个真实的 LBSN 数据集上兴趣点预测的 PR 曲线,图 12 展示了 TGE-LBSN 算法

的兴趣点预测的 ROC 曲线展示.通过四个数据集上的对比,可以发现相对于好友预测 TGE-LBSN 算法在四个数据集上预测精确度差异并不大,算法在四个数据集上进行兴趣点预测性能更加稳定.

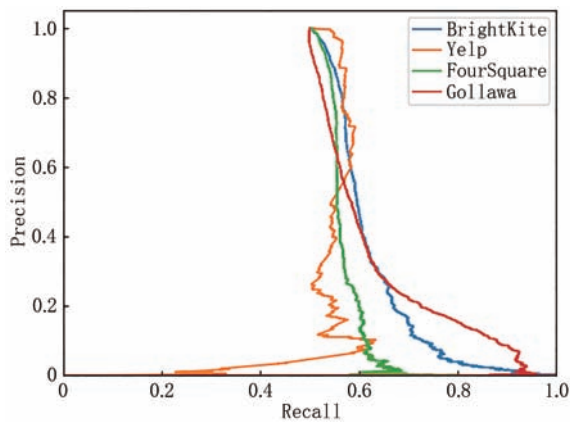


图 11 兴趣点预测 PR 曲线图

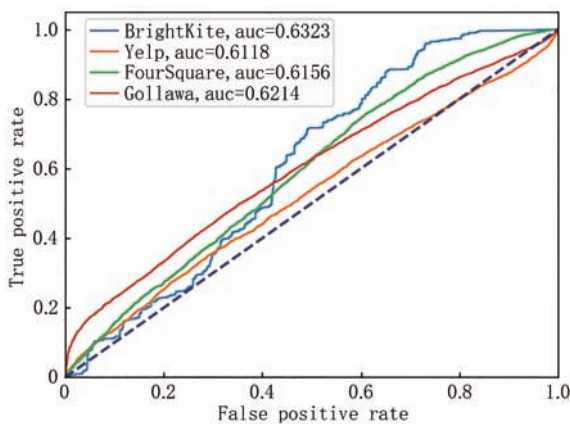


图 12 兴趣点预测 ROC 曲线图

#### 6.4 消融测试

为了验证 TGE-LBSN 算法所提创新点对于获得异构网络节点向量表示具有有效性,实验部分还进行了消融测试,并依据本文所提创新点,设计了四处测试,分别是:

(1) 时序划分(No-Time):本文通过删除阶段二的时序划分,不对时间进行排序,对所有的数据进行统一的训练,以测试时序划分对于不同数据集的影响程度;

(2) 有偏采样(No-Biased sampling):本文在这部分实验过程中,针对 LBSN 数据集,在原始 LBSN 数据集随机抽取与 TGE-LBSN 实验相同数量的数据集,即对于用户节点的多条 Check-in 数据,采用随机选取的方式确定 Check-in 数据,而 Check-in 数据的数量采用等比例抽取,以测试有偏采样过程对于实验结果的影响.同时在时序社交网络上同样采用随机选取等同数量的用户之间好友数据,进行验证有偏采样策略的有效性.

(3) 阶段一的初始向量(No-Initialvector):实验过程中利用随机的初始 One-hot 向量代替阶段

一的初始向量作为阶段二的输入,以测试阶段一的预处理及初始向量对于阶段二的预测效果的影响程度.

(4) 选择聚合邻居策略(No-SAN):通过在邻居聚合过程,对聚合邻居采用随机抽取的方式来测试本文所提的选择聚合邻居策略 SAN 对于 TGE-LBSN 算法性能的影响.

本文的实验分别在 LBSN 的两个数据集上以及时序社交网络的两个数据集上进行好友预测,依据以上三个不同测试,与完整的算法 TGE-LBSN 进行对比.实验结果展示在图 13,在前两个时序社交数据集上,有偏采样策略对于其影响比较大,即选择有代表性的友谊信息相较于随机选取友谊信息还是对于实验的影响比较大.对于 LBSN 数据集来说,有偏采样的策略对于实验结果同样有影响,由图可知对于 LBSN 数据集来说,若随机选取用户的 Check-in 数据,便会选取到冗余信息,这并不利于发现用户的偏好,对应实验结果表现,相较于不采用有偏采样策略,实验结果在 Yelp 数据集上提升 15.97%,在 Gollawa 数据集上可提升 32.92%.其余的策略对于时序社交网络并未太大影响,但是在 LBSN 数据集上却显现出了明显的差异,这与 6.2 节的结果分析相吻合,TGE-LBSN 算法更适用于 LBSN 数据集的节点向量生成.

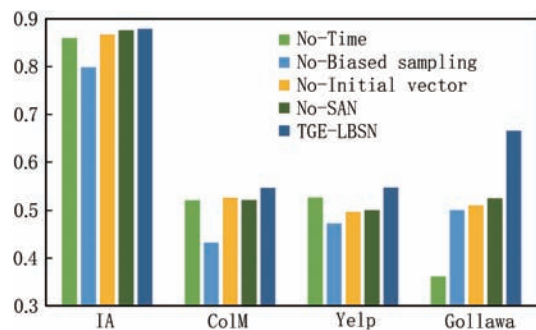


图 13 消融测试结果展示

综合分析发现,在 Yelp 数据集上,阶段一的初始向量生成的对性能影响比较大,这也就意味着即阶段一得到的节点原始特征向量对阶段二向量生成影响力很大,可以在一定程度上验证了对于 Yelp 的用户节点来说,阶段一聚合的自身特征、位置信息对于好友预测更有影响力.其次结果差异比较大的是邻居选择操作,选了更有表征性的邻居节点作为聚合选择,在实验过程中 TGE-LBSN 在训练时间上明显比 GraphSAGE、GAT 快,说明本文针对 LBSN



设计的SAN算法,不仅在一定程度上提高了训练速度,节省了训练时间,同时能提高学得节点向量的质量.由于实验结果不注重时间对比,所以此处并未给出数值化的时间对比.

在Gollawa数据集上,时序划分对预测结果影响最大,可见Gollawa数据集对时间因素比较敏感,这一定程度上体现了时序划分对其性能提升有一定的帮助,对于选择邻居策略以及阶段一初始向量生成,对于Gollawa来说同样很重要,说明这两种创新点为捕获Gollawa中有价值的信息做了一定的贡献,三种创新点在Gollawa数据集上更能体现TGE-LBSN算法在捕捉LBSN信息中具有高效性.

## 7 总 结

本文针对基于位置社交的图表示学习提出了TGE-LBSN表示学习框架,通过对LBSN进行有效的表示学习,确保节点的自身特征属性以及异构时空网络的结构信息和语义信息能被保留在学得节点向量表示中.首先依据有偏采样后,得到初始向量生成,并设计SAN算法来选取聚合的邻居,借助SVG算法得到子图下的节点向量表示.最后,借助注意力机制学得各个子图的自适应权重,将子图下的特征向量进行汇合得到最终节点的向量表示.在大量的真实LBSN数据集上的实验表明,本文提出的TGE-LBSN框架能有效提取LBSN中蕴含的丰富信息并得到有效的节点向量表示,在社交领域的好友推荐、兴趣点推荐等任务上均优于代表性的基准方法.

在未来的工作中,将会从以下方向进行深入研究:(1)将在LBSN上的表示学习拓展到更多社交领域预测任务,并依据不同的任务对学得的向量进行反向调控;(2)依据LBSN的隐私保护的现实需求,结合异地数据的联邦学习思想,未来考虑在LBSN上探索联邦学习角度下的异构网络表示学习算法.

## 参 考 文 献

- [1] Ren Xing-Yi, Song Mei-Na, Song Jun-De. Point recommendation of the contextual perception of location social network. *Journal of Computers*, 2017, 40 (4): 824-841 (in Chinese)  
(任星怡, 宋美娜, 宋俊德. 基于位置社交网络的上下文感知的兴趣点推荐. *计算机学报*, 2017, 40(4):824-841)
- [2] Tao Yong-Cai, Li Jun-Yan, Shi Lei, Wei Lin. Hybrid Location-based Personalized News Recommendation. *Journal of Chinese Computer Systems*, 2016, 37 (05): 943-947 (in Chinese)  
(陶勇才, 李俊艳, 石磊, 卫琳. 基于地理位置的个性化新闻混合推荐研究. *小型微型计算机系统*, 2016, 37(05):943-947)
- [3] Gong Wei-Hua, Shen Song, Pei Xiao-Bin, Yang Xu-Hua. Clustering and association method of dual heterogeneous communities in location-based social networks. *Chinese Journal of Computers*, 2020, 43(10):1909-1923 (in Chinese)  
(龚卫华, 沈松, 裴小兵, 杨旭华. 基于位置的社交网络中双重异质社区的聚类与关联方法. *计算机学报*, 2020, 43(10):1909-1923)
- [4] Ye M. Yin P. Lee W C. Lee D L. Exploiting geographical influence for collaborative point-of-interest recommendation// *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, China, 2011:325-334
- [5] Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y. GeoMF: Joint Geographical Modeling and Matrix Factorization for Point-of-Interest Recommendation// *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, USA, 2014:831-840
- [6] Wang Ke-Qiang, Peng Hong-Wei, Jin Yuan-Yuan. Sha Chao-Feng, Wang Xiao-Ling. Local Weighted Matrix Factorization for Top-n Recommendation with Implicit Feedback. *Data Science and Engineer*, 2016, 1(4):252-264
- [7] Yang Ding-Qi, Qu Bing-Qing, Yang Jie, Cuder-Mauroux P. LBSN2Vec++ : Heterogeneous Hypergraph Embedding for Location-Based Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020, PP (99):1-1
- [8] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations// *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA, 2014: 701-710
- [9] Bonner S, Kureshi I, Brennan J, Theodoropoulos G, McGough A S, Obara B. Exploring the Semantic Content of Unsupervised Graph Embeddings: An Empirical Study. *Data Science and Engineer*, 2019, 4(3):269-289
- [10] Yuan Quan, Cong Gao, Ma Zong-Yang, Sun Ai-Xin, Magnenat-Thalmann N. Time-aware point-of-interest recommendation// *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland, 2013: 363-372
- [11] Cao Shao-Sheng, Lu Wei, Xu Qiong-Kai. Grarep: Learning graph representations with global structural information// *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. Melbourne, Australia, 2015: 891-900
- [12] Wu Shi-Wen., Zhang Yuan-Xing, Gao Cheng-Liang, Bian Kai-Gui. GARG: Anonymous Recommendation of Point-of-Interest in Mobile Networks by Graph Convolution Network. *Data Science and Engineer*. 2020, 5(4):433-447
- [13] Ma Jing-Wei., Wen Jia-Hui., Zhong Ming-Yang, Chen Wei-

- Tong, Li Xue. MMM: Multi-source Multi-netMicrovideo Recommendation with Clustered Hidden Item Representation Learning. *Data Science and Engineer*, 2019, 4(3):240-253
- [14] Lecun Y, Bengio Y, Hinton G. Deep learning. *nature*, 2015, 521(7553): 436-444
- [15] Qu Ming-Dong, Cui Peng, Pei Jian, Zhang Zi-Wei, Zhu Wen-Wu. Asymmetric Transitivity Preserving Graph Embedding// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016;1105-1114
- [16] QiuJie-Zhong, Dong Yu-Xiao, Ma Hao, Li Jian, Wang Kuan-San, Tang Jie. Network embedding as matrix factorization: Unifying DeepWalk, Line, PTE, and node2vec// *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. Marina Del Rey, USA, 2018; 459-467
- [17] Tang Jian, Qu Meng, Wang Ming-Zhe, Zhang Ming, Yan Jun, Mei Qiao-Zhu. Line: Large-scale informationnetwork embedding// *Proceedings of the 24th Inter-National Conference on World Wide Web*. Florence, Italy, 2015; 1067-1077
- [18] Grover A, Leskovec J. node2vec: Scalable feature learning for networks// *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016; 855-864
- [19] Tsitsulin A, Mottion D, Karras P, Müller E. Verse: Versatile graph embeddings from similarity measures// *Proceedings of the 2018 World Wide Web Conference*. Lyon, France, 2018; 539-548
- [20] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks// *Proceedings of the 5th International Conference on Learning Representations 2017*, Toulon, France, 2017;1-14
- [21] Jin Wei, Derr T, Wang Yi-Qi, Ma Y, Liu Zi-Tao, Tang Ji-Liang. Node Similarity Preserving Graph Convolutional Networks// *Proceedings of the WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining*. Virtual Event, Israel, 2021;148-156
- [22] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graph// *Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Long Beach, USA, 2017;1024-1034
- [23] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks// *Proceedings of the ICLR 6th International Conference on Learning Representations*. Vancouver, Canada, 2018;1-12
- [24] Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2012, 3(2): 1-159
- [25] Li Zi-Jian, Zheng Xue-Ling, LinXue-Lin, Zhao Zi-Yuan, et al. TransN: Heterogeneous Network Representation Learning by Translating Node Embeddings// *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering*. Dallas, USA, 2020;589-600
- [26] Dong Y, Chawla N Y, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks// *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, Canada, 2017; 135-144
- [27] Fu T Y, Lee W C, Zhen L. HIN2Vec: Explore Me-ta-paths in Heterogeneous Information Networks for Representation Learning// *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Singapore, 2017;1797-1806
- [28] GuiHuan, Liu Ji-Lu, TaoFao-Bo, Jiang Meng, Norick B, Han Jian-Wei. Large-Scale Embedding Learning in Heterogeneous Event Data// *Proceedings of the 2016 IEEE 16th International Conference on Data Mining*. Barcelona, Spain, 2017;907-912
- [29] Ke T, Peng C, Xiao W, et al. Structural Deep Embedding for Hyper-Networks// *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA 2018;426-433
- [30] HuZi-Niu, Dong Yu-Xiao, WangKuan-San. Heterogeneous Graph Transformer// *Proceedings of The Web Conference 2020*, Taipei, Taiwan, 2020;2704-2710
- [31] Xie Min, Yin Hong-Zhi, Wang Hao, Xu Fang-Jiang, Chen Wei-Tong, Wang Sen. Learning Graph-based POI Embedding for Location-based Recommendation// *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, USA, 2016;15-24
- [32] Yang Ding-Qi, Qu Bing-Qing, Yang Jie, Cudre-Mauroux P. Revisiting User Mobility and Social Relationships in LBSNs: A Hypergraph Embedding Approach// *The World Wide Web Conference Association for Computing Machinery*, New York, USA, 2019;2147 - 2157
- [33] Yu Bing, Yin Hao-Teng, Zhu Zhan-Xing. 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
- [34] WuZong-Han, Pan Shi-Rui, Long Guo-Dong, Jiang Jing, Zhang Cheng-Qi. Graph WaveNet for Deep Spatial-Temporal Graph Modeling// *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, 2019; 1907-1913
- [35] Schlichtkrull M, Kipf T N, Bloem P, Berg R, Titov I, Welling M. Modeling Relational Data with Graph Convolutional Networks// *Proceedings of the 2018 European Semantic Web Conference*. Crete, Greece, 2018;593-607
- [36] Sun Yi-Zhou, Han Jia-Wei, Yan Xi-Feng, Yu P S. Wu Tian-Yi. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks// *Proceedings of the 37th International Conference on Very Large Data Bases Endow*, Westin, USA, 2011;992-1003



**ZHAO Lin-Lin**, M. S. candidate.

Her research interests include graph neural network, located social network.

**WU An-Biao**, Ph. D. candidate. His research interests include graph database, graph neural network.

**YUAN Ye**, Ph. D., professor. His main research interests include cloud computing, big data management (including graph data management, uncertain data management, data privacy protection) and P2P computing.

**LI Yang**, M. S. candidate. His research interests include graph neural network, graph embedding.

**WANG Guo-Ren**, Ph. D., professor. His main research interests include uncertain data management, data-intensive computing, unstructured data management, distributed query processing and optimization technology.

## Background

Location-Based Social Networks has been extensively studied due it is widely applied in the area of user preference analysis, recommendation system and information diffusion analysis, etc. At the same time, Graph embedding shows good results by virtue of its powerful data fitting ability. According to different graph structures, graph representation learning can be divided into homogeneous network representation learning and heterogeneous network representation learning. In this article, we mainly use LBSN to use heterogeneous network representation learning to obtain its node vector.

This research belongs to the application of graph representation to location social networks. The current international related research is LBSN2vec++, which mainly uses the basic ideas of DeepWalk.

This paper proposes TGE-LBSN (Two Stages of Graph Embedding on LBSN) for LBSN-oriented two-stage graph representation learning, which transforms LBSN into a heterogeneous network, and automatically extracts the features of LBSN with the help of graph representation learning to obtain nodes' vector representation with sufficiently rich information. First, the LBSN Check-in super edges are biased sampled according to users' Check-in time. In the first stage, the IVGS (Initial Vector Generation Stage) algorithm is designed, the friendship edges and the Check-in super edges are used to jointly

generate nodes' vectors containing position and feature information by IVGS algorithm. The generated nodes' vectors are used as the input of the second stage. Second, the second stage is mainly responsible for generating the final nodes' vectors in LBSN. LBSN is divided into different subgraphs according to the Check-in time, and we design the LBSN-oriented SAN (Select Aggregated Neighbors) strategy which is used to select representative neighbors to complete the aggregation operation, and then use the subgraph vector generation algorithm SVG (Subgraph Vector Generation) to obtain the vector representation of the nodes in each subgraph. Finally, the loss function is set according to the downstream tasks, and the attention mechanism is also used to learn adaptive weights for the subgraphs in different time periods to obtain the final vectors of the nodes, and then we use the final nodes' vectors to complete various prediction tasks. In future work, according to the actual needs of LBSN privacy protection, we will focus on combining with the idea of federated learning of remote data, the future can realize the heterogeneous representation learning on LBSN from the perspective of federated learning.

This paper is supported by Foundation item: National Natural Science Foundation of China (61932004, 62002054, 61732003, 61729201, 61972077); Research Funds for the Central Universities (N181605012), and the Liaoning Revitalization Talents Program (Grant No. XLYC2007079).