

基于隐藏特征的增强GNNs跨网络身份关联算法

潘语泉, 袁得崙, 王安然, 贾源

引用本文

潘语泉, 袁得崙, 王安然, 贾源. 基于隐藏特征的增强GNNs跨网络身份关联算法[J]. 计算机科学, 2026, 53(8): 50-60.

PAN Yuquan, YUAN Deyu, WANG Anran, JIA Yuan. Enhanced GNNs Across Social Networks User Identity Linkage Algorithm Based on HiddenFeatures [J]. Computer Science, 2026, 53(8): 50-60.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[融合动态可组合多头注意力的深度学习侧信道分析方法研究](#)

Research on Deep Learning-based Side-channel Analysis Method with Dynamically Composable Multi-head Attention

计算机科学, 2026, 53(8): 437-445. <https://doi.org/10.11896/jsjcx.250600213>

[面向漏洞检测的代码表征学习研究综述](#)

Survey on Code Representation Learning for Vulnerability Detection

计算机科学, 2026, 53(8): 388-402. <https://doi.org/10.11896/jsjcx.250900036>

[面向药物-靶点相互作用预测的门控双向Mamba多模态特征融合框架](#)

Gated Bidirectional Mamba Multimodal Feature Fusion Framework for Drug-Target Interaction Prediction

计算机科学, 2026, 53(8): 326-335. <https://doi.org/10.11896/jsjcx.250500096>

[CCSFR:基于协同-内容语义融合的评论增强推荐模型](#)

CCSFR: Collaborative-Content Semantic Fusion for Review-enhanced Recommendation

计算机科学, 2026, 53(8): 276-284. <https://doi.org/10.11896/jsjcx.250800027>

[多源关联知识图谱增强的图网络交通预测模型](#)

Leveraging Multi-source Contextual Knowledge-enhanced Graph for Traffic Forecasting

计算机科学, 2026, 53(8): 245-256. <https://doi.org/10.11896/jsjcx.250900014>

基于隐藏特征的增强 GNNs 跨网络身份关联算法

潘语泉¹ 袁得崙^{1,2} 王安然¹ 贾源¹

1 中国人民公安大学信息网络安全学院 北京 100038

2 安全防范与风险评估公安部重点实验室 北京 102623

(1870722711@qq.com)

摘要 跨网络身份关联能够判别来自不同社交网络的虚拟用户是否属于同一自然人。为应对真实数据集中正负样本分布的均衡程度对用户身份判别准确性的影响,提出了基于隐藏特征的增强 GNNs 跨网络身份关联算法。首先,为充分利用社交网络结构信息,挖掘了其中的隐藏特征,明确了度较大的节点和一阶邻居节点的重要作用;其次,设计了 w-LINE 算法,用于生成结构特征向量,通过与结构特征值向量进行拼接,得到用户特征向量;再次,提出了增强 GNNs,其由 Feature-GCN, Dynamic-Weight-GAT, GAE-VAE 组成,用于优化用户向量表达;最后,引入了自适应胶囊网络,自定义路由层和损失权重的设置,能够灵活应对不同的正负样本分布,实现跨网络身份关联。在两个真实数据集上开展对比实验,结果表明,与基线模型相比,所提算法在 P, R, F1 值中均有 10% 以上的提升。消融实验结果表明,各组件对算法的整体性能具有重要作用。

关键词: 跨社交网络;身份关联;图神经网络;胶囊网络;深度学习

中图分类号 TP391

Enhanced GNNs Across Social Networks User Identity Linkage Algorithm Based on Hidden Features

PAN Yuquan¹, YUAN Deyu^{1,2}, WANG Anran¹ and JIA Yuan¹

1 School of Information Network Security, People's Security University of China, Beijing 100038, China

2 Key Laboratory of Security and Risk Assessment, Ministry of Public Security, Beijing 102623, China

Abstract Across social networks user identification can determine whether virtual users from different social networks correspond to the same natural person. In order to address the influence of the equilibrium degree of the distribution of positive and negative samples in real datasets on the accuracy of user identity discrimination, an enhanced GNNs cross-network identity association algorithm based on hidden features is proposed. Firstly, in order to make full use of the structural information of social networks, the hidden features therein are mined, and it is clarified that nodes with larger degrees and first-order neighbor nodes play important roles. Secondly, the w-LINE algorithm is designed to generate the structural feature vectors. By concatenating them with the structural feature value vectors, the user feature vectors are obtained. Then, enhanced GNNs are proposed, including Feature-GCN, Dynamic-Weight-GAT, and GAE-VAE, for optimizing the expression of user vectors. Finally, the adaptive capsule network is introduced. With the settings of the custom routing layer and loss weights, it can flexibly cope with different positive and negative sample distributions and achieve cross-network identity association. Experiments are conducted on two real datasets. The comparative experimental results with the baseline model show that the proposed algorithm improves by more than 10% in the values of P, R, and F1. The results of the ablation experiment show that each component plays an important role in the overall performance of the proposed algorithm.

Keywords Across social networks, User identity linkage, Graph neural network, Capsule network, Deep learning

1 引言

随着数字时代的到来,社交平台迅速发展,为用户提供了广泛的自我表达与互动空间。然而,这也引发了一系列

新的挑战,尤其是如何在不同社交平台之间准确地识别并匹配同一用户身份^[1]。这一问题不仅是社交网络研究领域的重要课题,也是构建精准用户画像和进行用户行为分析的基础。在这样的背景下,跨网络身份关联对构建全面的用户画像至

到稿日期:2025-05-19 返修日期:2025-07-18

基金项目:公安部技术研究计划重点项目(2024JSZ01)

This work was supported by the Key Project of the Ministry of Public Security's Technical Research Program(2024JSZ01).

通信作者:袁得崙(yuandeyu@ppsuc.edu.cn)

关重要。它允许将分散于不同在线平台上的用户数据片段整合起来,形成一个更加完整、立体的用户形象^[2]。由于用户可能在各平台上使用不同的账户名称、头像等标识符,加之个人隐私保护意识的提升,实现用户身份的关联变得愈加困难。如何在确保个人信息安全的前提下,有效收集和使用数据,是解决问题的关键所在。

传统的用户身份关联方法主要依赖于用户提供的直接信息,如用户发布的信息或用户签到信息等^[3],但这些方法往往受到数据可获取性和隐私政策的限制。近年来,机器学习与深度学习技术的进步,特别是图神经网络(Graph Neural Networks, GNNs)的应用,为解决用户身份关联的问题提供了新的途径。图神经网络能够高效处理和利用网络中的结构化信息,这对于理解和解析用户之间的关系及其行为模式具有显著作用^[4-5]。通过构建和分析社交网络关系, GNNs 可以发现潜在的连接模式,即便在缺乏明确标识信息的情况下,也能有效促进用户身份的识别与关联。尽管 GNNs 已经在处理网络结构化数据方面表现出强大的能力,但其应用仍面临若干挑战。传统 GNNs 往往侧重于捕捉节点间的直接连接和局部邻域信息,而忽略了隐藏在复杂网络结构中的深层次特征。例如,度较大的节点和一阶邻居节点所蕴含的重要信息对用户身份关联具有显著意义,但在实际操作中,这些特征可能未被充分挖掘。此外,当前社交网络的结构呈现出多样化的节点度分布模式,不同平台上的好友关系网络在拓扑结构上存在显著差异,正样本(已确认为同一用户的跨平台用户对)与负样本(未确认为同一用户的用户对)之间的数量分布也会存在严重不平衡的问题。现有方法在应对上述挑战方面仍存在一定局限性,从而影响了跨平台用户身份识别的准确性。

基于上述分析,为了充分利用社交网络结构信息,有效应对正负样本分布多样化(包括分布均衡或不均衡)的社交网络场景,提高对不同社交网络中虚拟用户同一性的判别能力,本文提出了基于隐藏特征的增强 GNNs 跨网络身份关联算法。本文的主要贡献如下:

1)挖掘了社交网络中的隐藏特征,证明了度较大的节点和一阶邻居节点对用户身份关联具有重要的作用。基于对隐藏特征的分析,科学选择了社交网络中的结构特征,并据此构建了结构特征值向量。

2)提出了 w-LINE 算法,它基于节点的度数调整边权重,并通过负采样和一阶及二阶近似方法最大化相邻节点和邻域内节点的共现概率,从而捕捉到更丰富的网络结构信息,生成用户的结构特征向量。将结构特征值向量与结构特征向量进行拼接得到长向量,形成了全新的用户特征表达方式。

3)构建了增强 GNNs(多层增强 GNN 架构),其由 Feature-GCN, Dynamic-Weight-GAT, GAE-VAE 组成,能够有效实现向量优化的功能。

4)提出了自适应胶囊网络,在不同数据集上高效完成了二分类任务,提升了算法的准确度。

2 相关工作

目前关于跨网络用户身份关联方面的研究,主要从用户

属性信息、用户行为信息、网络拓扑结构、深度学习 4 个方面展开。基于用户属性信息的方法利用静态和动态资料进行身份识别,但面临数据易变性和隐私问题。基于行为信息的方法通过分析签到、发帖等多维度行为数据捕捉用户模式,但性能在数据稀疏和动态变化时受限。基于网络拓扑结构的方法依赖社交关系图谱匹配用户,对用户特征提取效果不佳。基于深度学习的方法使用图神经网络生成节点嵌入,能自动提取复杂特征,但依赖标注数据且计算成本高。各方法在准确性、数据需求、适用场景等方面各有优劣,未来需融合多种方法以提升综合性能。

2.1 基于用户属性信息的身份关联方法

用户属性信息涵盖了个人资料中的静态和动态数据,如用户名、头像、兴趣爱好等,同时也包括用户生成的内容,这些都能反映了用户的偏好、行为习惯和社会关系。

Liu 等^[6]提出一种跨平台用户身份识别方法,将用户数据分为个人信息、语义模式和关键词模式 3 类,并结合反向传播算法学习时间偏移权重,构建语义与关键词序列相似度计算模型,提升身份匹配准确性。Zhang 等^[7]提出结合偏好逻辑分析与机器学习的方法,通过构建用户兴趣与行为模型,利用逻辑回归对比种子节点及其邻域信息,逐步扩展匹配结果,实现跨社交网络的用户身份识别。Liu 等^[8]提出一种基于用户名相似度传播的用户身份关联方法,针对中文社交网络特点,分离中英文字符并计算最大公共子串,结合好友网络结构,利用一级邻居信息评估用户匹配度。Anisa 等^[9]研究了社交网络中基于强标识符(如用户名)和弱标识符(如兴趣)的用户匹配问题,提出结合机器学习与优化算法的框架,并在多个数据集上验证其有效性。Li 等^[10]提出一种仅基于用户显示名称的监督学习方法,通过分析用户命名模式构建特征向量,并利用机器学习判断跨平台名称是否属于同一用户。Ding 等^[11]基于用户多平台属性特征,构建朴素贝叶斯模型,通过已匹配用户对的属性相似度预测潜在匹配,实现跨平台身份关联。Qu 等^[12]提出基于友谊网络的跨社交网络用户识别算法 FNUI,通过分析多跳好友相似度并用梯度下降优化权重,融合多重属性信息,结合 Gale-Shapley 算法实现精准用户匹配。Li 等^[13]提出一种基于显示名称的低成本跨网用户识别方法,通过提取命名模式中的信息冗余特征并结合机器学习,实现身份匹配。

然而,用户属性(如用户名、头像和兴趣爱好等)容易随时间变化,这可能导致匹配结果不稳定。此外,用户生成的内容和行为模式也可能随时间和情境变化,影响长期跟踪的准确性。

2.2 基于用户行为信息的身份关联方法

用户行为信息包括用户的签到地点、发帖时间、社交互动等。通过分析这些多维度的行为数据,可以捕捉用户的独特行为模式,从而实现更精准的跨平台身份识别。

Dai 等^[14]通过网格聚类筛选潜在相关签到数据,提取时空特征并优化多属性相似度权重,综合计算用户匹配得分,有效提升在签到数据失衡情况下的身份关联准确率。Huo^[15]提出了一种基于深度内容语义相关性的用户身份关联方法和

一种融合行为轨迹与内容语义的用户身份关联方法,通过引入统计和深度语义特征提取以及行为数据质量评估等模块,提高了跨社交媒体用户身份关联的效果。Li 等^[16]基于签到轨迹,提出一种结合时空耦合的用户识别方法,利用核密度估计、权重分配与不一致性惩罚提升跨账户匹配准确率,验证了其在身份关联中的有效性。Zheng 等^[17]提出了 JORA 模型,通过图卷积网络学习各社交网络中用户的行为信息,并联合优化表示与对齐过程,使不同网络中的用户映射到统一空间,自适应地学习跨网用户相似性。文献[18-19]基于用户在位置社交网络中的时空签到行为,提出了针对性完成 UIL 的方法。该方法结合停留区域划分、连续时间相似度计算和位置频繁模式模型,有效提升了跨社交网络身份关联准确率。Hang 等^[20]提出 UniRank 框架,通过融合局部与全局特征,结合全局与局部匹配策略,提升跨社交网络身份关联的准确率与覆盖率。Kong 等^[21]提出了 MNA 模型,从用户推文中提取空间、时间和文本特征,并融合社交特征计算账户间的相似性,进而采用稳定匹配策略完成用户身份识别。Ma 等^[22]提出 CP-Link 方法,基于用户时空签到行为,结合停留区域划分、改进连续时间相似度 (IDTW) 和位置频繁模式 (LFP) 模型,有效提升跨社交网络身份关联准确率。

基于用户行为信息的身份关联方法面临数据依赖性强、动态行为建模不足、多模态融合有限及标注依赖等问题,影响其在稀疏、跨网络场景下的性能与适用性。

2.3 基于网络拓扑结构的身份关联方法

网络拓扑结构主要指用户的社交关系,体现了用户之间的连接。每个用户被视为图中的一个节点,而朋友关系则通过边来表示。这种结构有助于识别不同社交平台上的共同好友圈,从而帮助确认同一用户在多个平台上的账号。

Liu 等^[23]提出 WL-Align 框架,通过扩展 Weisfeiler-Lehman 同构测试,结合跨网络标签更新与邻近性保持的表示学习,生成具有区分性的节点嵌入,实现网络间的精准关联。Chen 等^[24]提出融合局部网络与超图结构的跨平台身份关联框架,结合多级图卷积与空间协调机制,在无需用户内容或人口信息的情况下,提升大规模社交网络中的匹配准确率。Chu 等^[25]提出 CrossMNA 方法,针对多网络场景生成跨网络 (Inter-vector) 与内部网络 (Intra-vector) 嵌入向量,有效整合多种结构信息,支持网络关联与下游分析任务。Man 等^[26]提出的 PALE 结合了网络嵌入与监督信息,学习稳定跨网络映射,生成反映结构特征的节点嵌入,用于预测用户在不同网络中的对应关系。Keiha 等^[27]提出一种新颖的身份关联框架,通过自动提取结构特征,在社交网络中实现了更准确的节点关系预测,有效提升了身份关联的效果。Chen 等^[28]提出名为 HFUL 的混合框架,通过开发全面的索引结构和新颖的剪枝策略来提高用户帐户链接的效率,并采用基于核密度估计的方法缓解数据稀疏性,以提升有效性。Zhou 等^[29]提出无监督模型 NWUIL,通过图神经网络将节点映射为 Wasserstein 空间中的分布,并将身份关联问题建模为最优网络传输问题,实现社交网络间的用户对齐。

基于网络拓扑结构的身份关联方法虽有一定成效,但社

交网络的动态性和复杂性可能导致拓扑结构频繁变化,影响匹配的稳定性;当用户的社交圈较小或较为封闭时,可用的拓扑信息有限,降低了身份关联的准确性。

2.4 基于深度学习的身份关联方法

深度学习能够生成网络节点的嵌入表示,揭示节点间的连接关系和隐藏特征,从而大幅提升跨平台用户身份识别的准确性。

Xiong 等^[30]提出了一种基于深度神经网络的用户身份链接新方法 DSANE,该方法首先通过完全过滤方法优化网络,然后将局部和全局信息嵌入节点的向量中,从而极大地丰富了节点的表示。Li 等^[31]提出 MFLink 方案,通过特征嵌入、图神经网络与注意力机制的多模态融合及对抗性学习,整合用户属性、内容和社会关系,增强跨平台用户表示相似性,有效解决数据语义差距并提升身份关联准确性。Huang 等^[32]提出基于双向图卷积网络 (B-GCN) 和多网络特征融合的模型 (MNF-UI),通过建模用户关系与提取跨社交网络的结构特征,实现多平台身份关联。Zhou 等^[33]提出的 DeepLink 是一种基于深度学习的用户身份链接方法,通过网络采样生成训练数据,将节点编码为向量并输入神经网络,结合双学习机制与策略梯度优化,实现跨社交网络的半监督用户匹配。Tang 等^[34]提出 CPUM 框架,基于图神经网络 AdaGAT 编码器,在联合向量空间中融合用户行为与属性信息,通过拓扑与属性一致性建模及位置编码,提升跨社交网络身份关联的准确性。Feng 等^[35]提出端到端的深度学习框架 DPLink,结合位置与轨迹编码器及共注意力机制,从异构移动数据中学习用户行为表征,实现跨服务的用户身份链接。Zhang 等^[36]提出的 Mego2Vec 是一种图神经网络模型,通过将用户自我中心网络映射为低维向量,融合多视图嵌入、统一注意力与结构嵌入,实现跨社交网络的用户关联。Shao 等^[37]提出 AsyLink 框架,通过主题建模关联文本与位置信息,构建用户交互张量并采用三维卷积神经网络提取匹配模式,结合排序损失解决数据不平衡问题,实现跨网络用户身份链接。

深度学习模型通常需要大量的标注数据来进行训练,这些数据在实际应用中可能难以获取;其次,模型的复杂性和计算成本较高,对硬件资源有较高的要求。

在分析现有技术局限性的基础上,本文提出了一种基于隐藏特征的增强 GNNs 跨网络身份关联算法,旨在弥补上述研究方法中的不足。针对依赖用户属性信息进行身份关联时遇到的数据获取难和隐私保护问题,本文算法通过挖掘社交网络中度数较大的节点及一阶邻居节点等隐藏特征来构建用户特征向量,减少了对直接用户属性信息的依赖。为解决基于用户行为信息的身份关联过程中特征提取困难的问题,本文提出的 w-LINE 通过调整边权重反映节点度数,并采用负采样与一阶、二阶近似策略,最大化邻近及邻域节点的共现概率,从而更全面地捕捉网络结构特征。面对网络拓扑结构分析中社交圈较小或封闭导致匹配精度降低的问题,设计了增强 GNNs,包括 Feature-GCN, Dynamic-Weight-GAT 和 GAE-VAE 3 个模块。增强 GNNs 能够有效优化用户向量表达,增强了模型处理复杂网络结构的能力。对于深度学习方法中常

见的标注数据需求大和计算成本高的挑战,本文引入了自适应胶囊网络,自定义路由层和损失权重的设置能够灵活应对不同的正负样本分布,提高了算法的泛化能力和识别准确率。通过上述创新点,本文算法有效解决了当前跨网络身份关联领域中存在的主要问题,并展示了其在不同数据集上的优越性能。

3 问题分析

3.1 相关定义

定义 1(社交网络) 社交网络(SN)可以被定义为一个由节点(代表个体或实体)和边(表示个体间的某种类型的关系)组成的图结构。对于 $SN=(V, E)$, V 是节点集,每个节点 $v \in V$ 代表社交网络中的一个成员, $E \subseteq V \times V$ 是边集,每条边 $e \in E$ 表示两个节点之间的直接联系。

定义 2(好友关系) 社交网络中,好友关系(FR)特指两个成员之间存在的一种特定类型的连接,通常意味着双方相互认可对方为朋友。对于任意两个节点 $u, v \in V$,如果它们之间存在一条边 $e_{uv} \in E$,则 $FR(u, v)=1$,否则 $FR(u, v)=0$ 。

定义 3(锚链接) 锚链接(AL)是指跨不同社交平台之间相同用户的身份标识链接,可帮助识别出在多个平台上活跃的同一个实际用户。对于两个社交网络 $SN_1(V_1, E_1)$ 和 $SN_2(V_2, E_2)$,如果存在一对节点 $u_1 \in V_1$ 和 $u_2 \in V_2$,分别属于两个不同的社交网络且对应同一个自然人,则称这对节点之间存在锚链接,且 $AL(u_1, u_2)=1$,否则 $AL(u_1, u_2)=0$ 。

定义 4(锚链接用户) 锚链接用户(ALU)指通过至少一条有效锚链接与另一个社交网络中的对应用户相连接的用户。对于一个社交网络 $SN_i(V_i, E_i)$,如果存在另一个社交网络 $SN_j(V_j, E_j)$,且 $i \neq j$,使得对于某个节点 $u_i \in V_i$,存在一个 $u_j \in V_j$,使得 $AL(u_i, u_j)=1$,则称 u_i 为锚链接用户,记作 $ALU(u_i)=1$,否则 $ALU(u_i)=0$ 。

定义 5(身份关联) 为了判断来自两个不同社交网络的虚拟用户是否属于同一自然人,可以将身份关联(UIL)任务转换为二分类问题。对于来自不同社交网络平台的 u_i 和 u_j ,有如下二分类结果:

$$UIL(u_i, u_j) = \begin{cases} 1, & \text{if } ALU(u_i) \cdot ALU(u_j) = 1 \\ 0, & \text{if } ALU(u_i) \cdot ALU(u_j) = 0 \end{cases} \quad (1)$$

3.2 特征分析与算法框架

3.2.1 特征分析

本文在针对身份关联的研究中,创新性地挖掘了在传统结构关系中的隐藏特征,并选择了相应的结构特征,作为构造用户特征向量的重要基础。

1) 隐藏特征分析

给定两个社交网络 $G_1=(V_1, E_1)$ 和 $G_2=(V_2, E_2)$,其中, V_1 和 V_2 分别是两个网络中的节点集合, E_1 和 E_2 分别是边的集合。

(1) 度排名较大的节点分析

根据每个网络中节点的度(即连接数),对节点进行降序排列。对于网络 G_1 ,得到一个有序列表 L_1 ;对于网络 G_2 ,得到一个有序列表 L_2 。将这两个列表各自分为 k 个大小相等或几

乎相等的子集(如果不能整除,则最后一个子集可能较小)。对于 G_1 ,子集记作 $S_{1,1}, S_{1,2}, \dots, S_{1,k}$;对于 G_2 ,子集记作 $S_{2,1}, S_{2,2}, \dots, S_{2,k}$ 。对于每一对对应的子集 $S_{1,i}$ 和 $S_{2,i}$ ($i=1, 2, \dots, k$),计算它们之间的交集大小,记作 $|S_{1,i} \cap S_{2,i}|$ 。

对于每一个 i ,定义重叠度 C_i 为:

$$C_i = \frac{|S_{1,i} \cap S_{2,i}|}{\min(|S_{1,i}|, |S_{2,i}|)} \quad (2)$$

(2) 一阶邻居节点分析

对于任意顶点 v 和整数 $l \geq 1$, $N^l(v)$ 表示从 v 出发,经过不超过 l 步可到达的顶点集合。对于任意顶点集合 V ,给定实体映射函数 $F, F(V)$ 表示 V 中所有顶点对应的实体集合。对于 G_1 中的顶点 v_i 和 G_2 中的顶点 v_j ,计算它们的 l -hop 邻居集合的邻域相似度,公式为:

$$s_l(v_i, v_j) = \frac{|F(N^l(v_i)) \cap F(N^l(v_j))|}{\min(|F(N^l(v_i))|, |F(N^l(v_j))|)} \quad (3)$$

2) 结构特征分析

考虑到节点度($D(v)$)能够直接反映一个节点在网络中的连接数量,是衡量节点中心性的重要指标之一;聚集系数($C(v)$)则用于评估节点的邻居间形成闭环的程度,进一步补充了关于节点局部网络密度的信息; $PageRank(PR(v))$ 作为另一个关键特征,不仅考虑了节点的直接连接数量,还考量了链接到该节点的其他节点的重要性,从而提供了关于节点全局重要性的信息;邻域度中心性($NDC(v)$)、邻居平均度($AND(v)$)、邻居平均聚集系数($ANCC(v)$)以及邻居平均($PageRank(ANPR(v))$)共同作用,有助于捕捉节点周边环境的结构特性。为了更加全面地理解节点与其邻居之间的相互关系,还加入了邻居平均邻域度中心性($ANND(v)$)。因此,基于隐藏特征分析强调的度较大节点和一阶邻居节点的重要性,结合上述分析,选择以上 8 个特征参与构造特征值向量。这 8 个特征综合反映了节点本身的属性及其所处的局部网络环境,为后续算法提供了丰富的输入信息,能够有效提高跨网络身份关联的准确性和鲁棒性。通过这样的方式,不仅强化了对用户特征向量表达的理解,也为后续实验验证奠定了坚实的理论基础。

上述特征的计算式如下。

(1) 节点度($D(v)$)

对于任意节点 V ,其度为与之直接相连的邻居数量:

$$D(v) = |\{u \in v | (v, u) \in E\}| \quad (4)$$

(2) 聚集系数($C(v)$)

节点 V 的聚集系数用于衡量它的邻居之间相互连接的程度。若 v 有 $d(v)$ 个邻居,则表示为:

$$C(v) = \frac{2e_v}{d(v)(d(v)-1)} \quad (5)$$

其中, e_v 表示 v 的邻居间实际存在的边的数量。

(3) $PageRank(PR(v))$

在一个图 G 中,节点 v 的 $PageRank$ 值是由所有指向它的节点 u 的 $PageRank$ 值加权求和得到的:

$$PR(v) = (1-d) + d \sum_{u \in B_v} \frac{PR(u)}{L(u)} \quad (6)$$

其中, d 是阻尼因子, B_v 是所有有边指向 v 的节点集合, $L(u)$ 是节点 u 的出度。

(4) 邻域度中心性($NDC(v)$)

邻域度中心性是指节点 v 的所有邻居的度之和,表示为:

$$NDC(v) = \sum_{u \in N(v)} D(u) \quad (7)$$

(5) 邻居平均度($AND(v)$)

节点 v 的邻居平均度是所有邻居节点度的算术平均数:

$$AND(v) = \frac{1}{|N(v)|} \sum_{u \in N(v)} D(u) \quad (8)$$

(6) 邻居平均聚集系数($ANCC(v)$)

节点 v 的邻居平均聚集系数是所有邻居节点聚集系数的算术平均数:

$$ANCC(v) = \frac{1}{|N(v)|} \sum_{u \in N(v)} C(u) \quad (9)$$

(7) 邻居平均 PageRank($ANPR(v)$)

节点 v 的邻居平均 PageRank 是所有邻居节点 PageRank 值的算术平均数:

$$ANPR(v) = \frac{1}{|N(v)|} \sum_{u \in N(v)} PR(u) \quad (10)$$

(8) 邻居平均邻域度中心性($ANND(v)$)

$$ANND(v) = \frac{1}{|N(v)|} \sum_{u \in N(v)} NDC(u) \quad (11)$$

3.2.2 算法框架

为判断来自两个社交网络的虚拟用户是否对应为同一自然人,本文提出了基于隐藏特征的增强 GNNs 跨网络身份关联算法,对于社交网络的朋友关系,基于隐藏特征分析,选择 8 个结构特征参与构造特征值向量;提出了 w-LINE 算法,得到用户的节点嵌入向量,将两者拼接,构造表征用户特征的长向量;设计了增强 GNNs,融合 Feature-GCN, Dynamic-Weight-GAT 和 GAE-VAE,实现了优化向量的功能;将优化后的用户特征向量作为自适应胶囊网络的输入,利用挤压函数与动态调节的耦合系数来处理输入信息,生成输出胶囊,并通过自适应二元交叉熵损失函数进行模型优化;经过训练和测试,输出最终的判别结果(是否为同一自然人)。算法基本框架如图 1 所示。

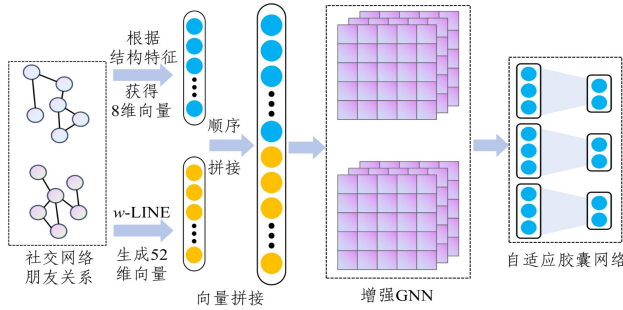


图 1 身份关联算法基本框架

Fig. 1 Basic framework of user identification algorithm

4 算法模型

4.1 w-LINE 算法

为了充分利用每个节点蕴含的隐藏特征,并生成相应的结构特征向量,提出了节点嵌入算法 w-LINE。其基本原理如下。

对于网络中任意一对节点 u 和 v ,其边的权重 w_{uv} 根据它们的度数 d_u 和 d_v 来计算。

$$w_{uv} = \frac{1}{(d_u + d_v)^\beta} \quad (12)$$

其中, β 是超参数,用于调整边权重与节点度数之间的关系。

在负采样过程中,首先需要构建一个基于节点度数的概率分布,每个节点 i 被选为负样本的概率 $p(i)$ 为:

$$p(i) = \frac{d_i^\alpha}{\sum_j d_j^\alpha} \quad (13)$$

其中, α 是超参数,用于抑制高次节点的影响。

在第一阶近似中,需要最大化相邻节点的共现概率,通过优化以下目标函数实现:

$$\max \left\{ \sum_{(u,v) \in E} \log \sigma(\mathbf{h}_u^T \cdot \boldsymbol{\theta}_v) + k \cdot E_{v' \sim p(v)} [\log \sigma(-\mathbf{h}_u^T \cdot \boldsymbol{\theta}_{v'})] \right\} \quad (14)$$

其中, $(u,v) \in E$ 表示网络中的边集合 E , u 和 v 是一条边的两个端点, $\mathbf{h}_u$ 是节点 u 的嵌入向量, $\boldsymbol{\theta}_v$ 是节点 v 的上下文嵌入向量, $\sigma(x)$ 是 sigmoid 函数, $\mathbf{h}_u^T \cdot \boldsymbol{\theta}_v$ 表示它们之间的相似度

得分, k 表示负采样的数量, $v' \sim p(v)$ 表示从负采样分布 $p(v)$ 中随机选择负样本节点 v' 。

在第二阶近似中,需要最大化节点与其邻域内其他节点的共现概率。每个节点有一个单独的上下文嵌入向量 $\boldsymbol{\theta}_v$ 。具体来说,节点 u 的嵌入向量可以看作其邻居上下文嵌入向量的加权平均:

$$s = \sum_{n \in N(u)} w_{un} \boldsymbol{\theta}_n \quad (15)$$

其中, $N(u)$ 表示节点 u 的邻居集合, w_{un} 是节点 u 和邻居 n 之间的边权重。

w-LINE 算法的原理如算法 1 所示。

算法 1 w-LINE 算法

Input: 社交网络用户关系信息

Output: 每个用户节点对应的结构特征向量

Begin

1. initialize dimensions, negative_samples, learning_rate, epochs, order, α , β .
2. add edges, 构建节点映射 node_map 与度分布 degree_distribution.
3. define 负采样函数 negative_sampling(node)
4. for epoch in epochs do
5. shuffle edges
6. for edge(u, v) in edges do
7. $w_{uv} \leftarrow \frac{1}{(d_u + d_v)^\beta}$
8. if order == 1 then
9. update embeddings for u and v ;

```

10.     grad ← σ(u_emb · v_emb) - 1
11.     u_emb ← u_emb - learning_rate · grad · v_emb · w_uv
12.     v_emb ← v_emb - learning_rate · grad · u_emb · w_uv
13.     else if order == 2 then
14.         updatecontext_embedding
15.     end if
16.     for each neg in neg_node do
17.         grad ← σ(u_emb · v_ctx) - 1
18.         u_emb ← u_emb - learning_rate · grad · v_ctx · w_uv
19.         v_ctx ← v_ctx - learning_rate · grad · u_emb · w_uv
20. return embedding
21. End
    
```

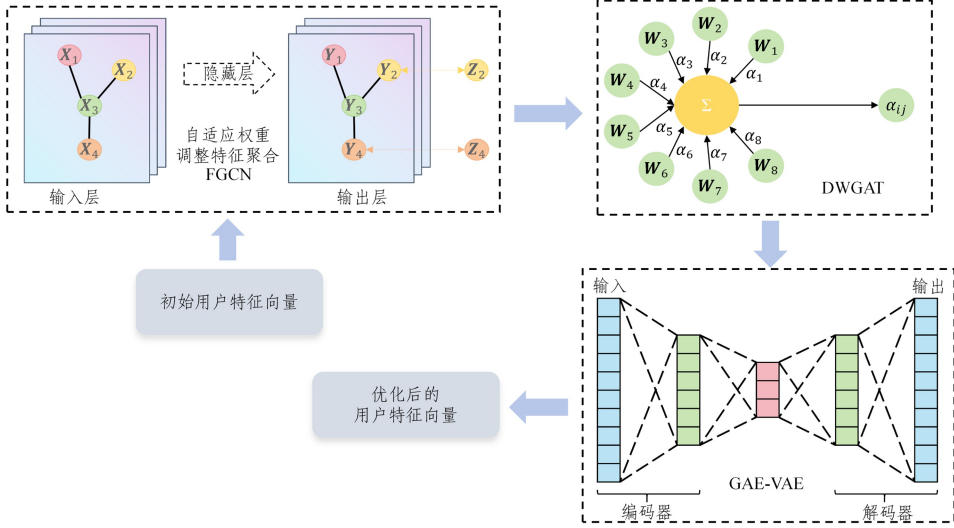


图2 增强 GNNs 架构

Fig. 2 Architecture of enhanced GNNs

1)FGCN

通过图卷积操作学习节点的隐藏表示,并对前 8 维特征应用自适应权重。图卷积的过程如式(16)所示:

$$H = \sigma(\tilde{A}XW) \quad (16)$$

其中, $\tilde{A}$ 是邻接矩阵的规范化形式,通常为 $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$; W 是可学习的权重矩阵; σ 是激活函数(如 ReLU)。特征加权的公式为:

$$H_{:,8} = H_{:,8} \odot W_f \quad (17)$$

其中, $H_{:,8}$ 表示 H 的前 8 个特征; W_f 是可学习的特征权重,通过重复扩展到与 $H_{:,8}$ 相同的维度; $\odot$ 表示逐元素乘法。

2)DWGAT

利用注意力机制动态调整邻接边的权重,强调邻居节点的重要性。通过多头注意力机制捕捉节点之间的不同关系,并通过线性变换进一步处理节点表示。

$$H_i' = \sigma\left(\sum_{k=1}^K \alpha_k \cdot \text{Att}_k(H, A)\right) \quad (18)$$

其中, K 是注意力头的数量, Att_k 是第 k 个注意力头的计算, α_k 是每个注意力头的权重。

在线性变换中:

$$H' = \sigma(H'W) \quad (19)$$

其中, W 是另一个可学习的权重矩阵。

3)GAE-VAE

通过变分自编码器(VAE)框架学习节点的低维嵌入表

4.2 增强 GNNs

为了增强对用户特征向量的优化功能,本文提出了增强 GNNs,包括 3 个模块,如图 2 所示。第一个模块是 Feature-GCN(FGCN),通过自适应权重调整特征聚合,结合图卷积操作整合邻居节点特征,生成并更新每个节点的特征向量,供下一个模块使用;第二个模块是 Dynamic-Weight-GAT(DWGAT),利用注意力机制动态调节边的权重,并通过多头注意力机制融合邻居节点的特征;第三个模块是 GAE-VAE,通过变分自编码器优化特征表示,结合重构误差与 KL 散度进行损失计算。

示,同时生成均值和方差,以便在重构图结构时引入不确定性,从而更好地处理数据中的噪声和变化。使用图卷积的计算式为:

$$H' = \sigma(\tilde{A}H'W) \quad (20)$$

在均值和对数方差计算中:

$$\mu = H'W_\mu \quad (21)$$

$$\log \sigma^2 = H'W_{\log \sigma^2} \quad (22)$$

其中, W_μ 是用于计算均值的权重矩阵, $W_{\log \sigma^2}$ 是用于计算对数方差的权重矩阵。

重参数化过程如下:

$$Z = \mu + \epsilon \odot \exp(0.5 \log \sigma^2) \quad (23)$$

其中, $\epsilon \sim \mathcal{N}(0, I)$ 是标准正态分布的噪声。

定义损失函数为:

$$L = L_{\text{recon}} + L_{\text{KL}} \quad (24)$$

其中,重构损失 L_{recon} 为:

$$L_{\text{recon}} = -\frac{1}{N^2} \sum_{i,j} [A_{i,j} \log \sigma(Z_i^T Z_j) + (1 - A_{i,j}) \log(1 - \log \sigma(Z_i^T Z_j))] \quad (25)$$

4.3 自适应胶囊网络

本文提出的自适应胶囊网络(ADCN)是一种新颖的神经网络架构,通过挤压函数和动态调整的耦合系数来处理输入数据,生成输出胶囊,并采用自适应二元交叉熵损失进行优化,能够根据正负样本的数量自动调整权重,强化识别能力。

输入 $\mathbf{X}$ 是一个形状为 $(N, 120)$ 的矩阵, 其中 N 是样本数量。对于输入向量 $\mathbf{v}$, 定义挤压函数:

$$\text{squash}(\mathbf{v}) = \frac{\|\mathbf{v}\|^2}{1 + \|\mathbf{v}\|^2} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|} \quad (26)$$

其中, $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^d v_i^2}$ 是向量 $\mathbf{v}$ 的欧几里德范数。

在自定义路由层中, 通过 $\mathbf{W}$ 变换后的输入胶囊向量 $\hat{\mathbf{u}}_{j|i}$ 的计算如下:

$$\hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij} \mathbf{u}_i \quad (27)$$

其中, $\mathbf{W}_{ij}$ 是从第 i 个输入胶囊到第 j 个输出胶囊的权重矩阵。

计算耦合系数时, 初始化 $b_{ij} = 0$, 然后通过 softmax 更新:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (28)$$

所有输入胶囊对第 j 个胶囊的加权和 s_j 为:

$$s_j = \sum_i c_{ij} \hat{\mathbf{u}}_{j|i} \quad (29)$$

应用挤压函数得到最终的输出胶囊 v_j :

$$v_j = \text{squash}(s_j) \quad (30)$$

通过多次迭代, 更新耦合系数 b_{ij} :

$$b_{ij} = b_{ij} + v_j \cdot \hat{\mathbf{u}}_{j|i} \quad (31)$$

假设正负样本比为 $a:b$, 使用自适应二元交叉熵损失函数 L 和 Adam 优化器训练模型。

$$L = -\frac{1}{N} \sum_{i=1}^N [\alpha y_i \log(\hat{y}_i) + \beta (1 - y_i) \log(1 - \hat{y}_i)] \quad (32)$$

其中, $\alpha = \frac{b}{a+b}, \beta = \frac{a}{a+b}$ 。

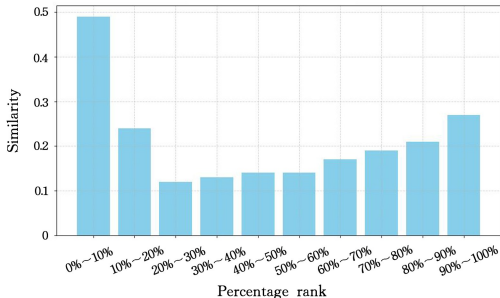
5 实验分析

5.1 数据集

为了充分模拟真实社交网络中的情况, 本文实验选择了 Twitter-Foursquare^[38] 和 Flickr-Lastfm^[39] 两组数据集, 它们都来源于真实的社交网络数据, 但正负样本的比例有较大差距。原始社交网络数据集中存在如下问题。

1) 异常用户的存在: 社交网络中存在一部分用户节点没有建立任何好友关系。这些孤立节点对于研究分析缺乏实质价值, 可能干扰研究结果的准确性。

2) 用户信息的不一致性: 因用户个人资料更新等因素, 数据库中某些用户之间的关联关系可能与现实情况不符。这种信息不一致会导致数据集中的节点信息与实际社交网络中的节点信息存在偏差, 影响数据的准确性和可靠性。



(a) Twitter-Foursquare

3) 单向朋友关系: 社交网络中存在单方面的好友关系, 即一方将另一方标记为好友, 而后者未将前者标记为好友。在构建社交网络拓扑图时, 若直接采用这些单向关系, 可能会增加网络结构的复杂度, 给分析带来困难。

因此, 对于每个社交网络, 剔除存在问题 1 和问题 2 的节点, 并删除相应的朋友关系。对于问题 3, 补齐所有的单向朋友关系, 均视为无权无向的边。数据集的信息如表 1 所列。

表 1 社交网络数据集信息

Table 1 Social network data set information

社交网络名称	用户节点数	用户关系数	锚链接数	负正样本比
Twitter	5 120	164 920	3 141	1:1.58
Foursquare	5 313	76 972		1:11.45
Flickr	173 078	1 515 133	452	381.92:11
Lastfm	136 409	1 685 469		300.40:11

5.2 评价指标

实验运行环境为 Ubuntu 22.04 操作系统, CPU 处理器为 Intel 酷睿 i7 13700K, 内存为 64GB DDR5 4800 MHz, 使用的软件为 Python 3.6 与 CUDA 12.1 兼容的 PyTorch、JupyterLab3。

采用传统的身份关联评价指标对模型性能进行评估, 评价指标包括精确率 P 、召回率 R 、综合评价参数 $F1$ 、准确率 Acc , 计算方法如下:

$$P = \frac{TP}{TP + FP} \quad (33)$$

$$R = \frac{TP}{TP + FN} \quad (34)$$

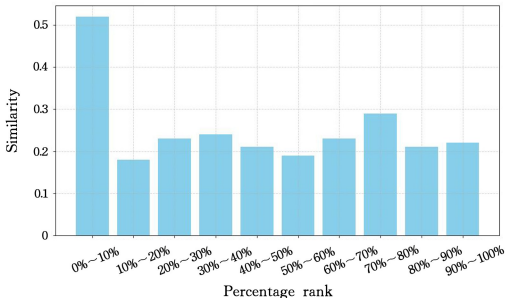
$$F1 = \frac{2PR}{P + R} \quad (35)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (36)$$

其中, TP 为正确预测为正样本的数量; FP 为错误预测为正样本的数量; TN 为正确预测为负样本的数量; FN 为错误预测为负样本的数量。

5.3 隐藏特征分析

对两组数据集分别进行隐藏特征分析, 结果如图 3—图 5 所示。图 4 和图 5 中, 纵轴表示邻居节点的相似度分布, 正方向表示节点属于同一实体(关联节点)的比例, 负方向表示节点不属于同一实体的比例。



(b) Flickr-Lastfm

图 3 度较大节点分析结果

Fig. 3 Analysis results of high-degree nodes

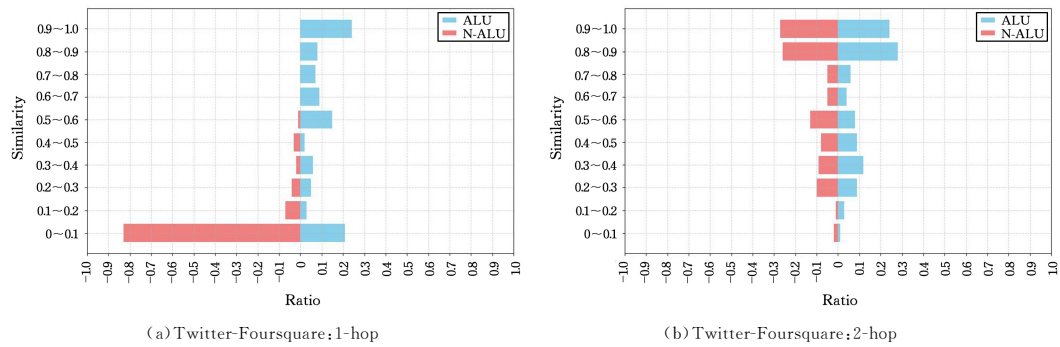


图 4 一、二阶邻居节点分析结果(Twitter-Foursquare)
Fig. 4 First and second-order neighbor node analysis results(Twitter-Foursquare)

对度较大节点的表明,同一实体在两组数据集中,对应节点的度排名具有一定的相似度。在度排名前 10%(即每张图的第一个柱状图)的节点集合中,两组数据集同一实体的重叠

度都在 0.5(前 10%的锚链接用户有 50%的节点对应同一实体),而其他度排名区间的节点集合相似度明显低于 0.5,说明度较大的节点对跨网络用户身份同一性的判别具有重要作用。

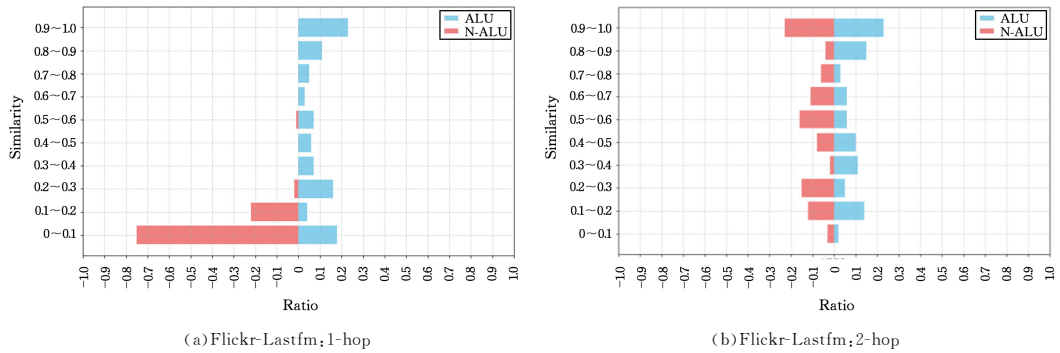


图 5 一、二阶邻居节点分析结果(Flickr-Lastfm)
Fig. 5 First and second-order neighbor node analysis results(Flickr-Lastfm)

对一、二阶邻居节点的分析结果表明:在 1-hop 的结果中,关联节点和非关联节点具有较强的区分度,且大部分关联节点的邻居相似度都在 0.8~1,对身份关联任务具有重要作用;在 2-hop 的结果中,两类节点没有较强的区分度,且对应相似度的分布较为零散,对身份关联的作用不明显。

居节点对身份关联任务具有重要作用,可基于此开展后续实验。

5.4 身份关联模型比较

除了本文模型,还选择了 WLAlign^[23],CrossMNA^[25],PALE^[26],B-GCN^[32],MNF-UI^[32]和 DeepLink^[33]在 Twitter-Foursquare 和 Flickr-Lastfm 两组数据集上分别进行对比实验,结果如表 2、图 6、图 7 所示。

综上所述,在社交网络数据集中,度较大的节点和一阶邻

表 2 身份关联模型对比实验结果						
Table 2 Comparative experimental results of user identification model						
身份关联模型	Twitter-Foursquare			Flickr-Lastfm		
	P	R	F1	P	R	F1
WLAlign	0.8087	0.8261	0.8173	0.7463	0.7291	0.7376
CrossMNA	0.7929	0.7845	0.7887	0.7359	0.7124	0.7240
PALE	0.7158	0.6992	0.7074	0.6753	0.7067	0.6906
B-GCN	0.8214	0.8036	0.8122	0.7259	0.7066	0.7158
MNF-UI	0.7856	0.7952	0.7906	0.6873	0.7129	0.6957
DeepLink	0.7334	0.7472	0.7402	0.7115	0.7348	0.7230
本文算法	0.9274	0.9011	0.9141	0.8137	0.8228	0.8182

由图表结果可知,本文算法在两个数据集上各项指标均有 10%以上的提升,充分说明了算法架构设计的合理性和科学性。CrossMNA 依赖于跨网络信息来优化两种节点的嵌入向量,即用于网络对齐的互向量和用于特定网络结构特征的内向量,但这种策略可能无法完全捕捉到所有网络中的结构差别。WLAlign 通过已知锚点的连通性来进行标签传播和

相似性哈希操作,在处理节点与标记锚点之间的长距离连接情况时仍然面临挑战。DeepLink 未能全面解决不同用户和社交平台异质性的问题,不同社交网络平台的独立运作及用户活动差异使得跨平台账户映射复杂。PALE 根据监督信息中的已观测锚点链接来进行网络嵌入和匹配,但对于缺乏此类监督信息的场景,尤其是正负样本不均衡时,其性能会受到

影响。B-GCN 依赖于精确计算用户间的亲密程度,并据此构建反映用户关系的邻接矩阵,因此对数据的质量与完整性提出了较高要求。MNF-UI 在处理多网络环境时,由于其对种子节点共同特征存在依赖,因此在实际应用中的灵活性及适用性受到一定限制;特别是在网络特征差异显著或种子节点信息不足的情况下,该方法的识别精度可能无法得到有效保证。

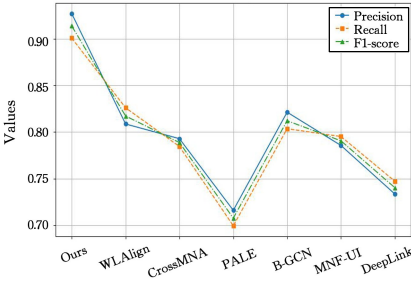


图 6 身份关联模型对比实验结果(Twitter-Foursquare)
Fig. 6 Comparative experimental results of user identification model (Twitter-Foursquare)

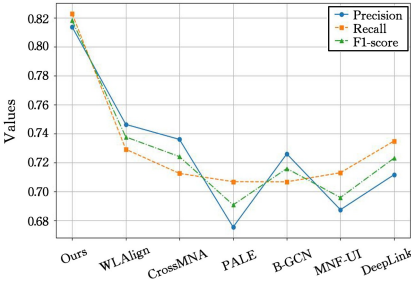


图 7 身份关联模型对比实验结果(Flickr-Lastfm)
Fig. 7 Comparative experimental results of user identification model (Flickr-Lastfm)

5.5 节点嵌入算法比较

LINE 优化邻近节点的共现概率来学习网络中节点的低维向量表示;Struc2vec 通过比较节点在网络中的结构角色,生成结构相似的节点嵌入;SDNE 结合自动编码器与图正则化,学习保留一阶和二阶邻近关系的节点嵌入。为比较各节点嵌入算法的性能,选择上述 3 种算法与本文提出的 w-LINE 进行节点嵌入效果的比较,如表 3 所列。

表 3 节点嵌入算法比较结果

节点嵌入算法	Twitter-Foursquare			Flickr-Lastfm		
	P	R	F1	P	R	F1
w-LINE	0.9274	0.9011	0.9141	0.8137	0.8228	0.8182
LINE	0.7142	0.7264	0.7202	0.6513	0.6457	0.6485
Struc2vec	0.6255	0.5917	0.6081	0.5882	0.6048	0.5964
SDNE	0.6838	0.6627	0.6731	0.6160	0.5925	0.6040

由表 3 可知,在两个数据集上,w-LINE 算法在 P, R 和 F1 分数方面均取得了最佳成绩,尤其是在 Twitter-Foursquare 数据集中,其 P 值达到了 0.9274,R 值为 0.9011,F1 得分为 0.9141,表明该算法能够高效地捕捉到节点间的共现概率,从而生成高质量的节点嵌入表示。相比之下,传统的 LINE 算法虽然也能较好地学习节点表示,但在所有 3 个评

价指标上都显著低于其加权版本 w-LINE,特别是在 Flickr-Lastfm 数据集上,其 P 值仅为 0.6513,R 值为 0.6457,F1 得分为 0.6485。这可能是因为标准 LINE 算法仅考虑了一阶邻近关系,而忽略了更复杂的网络结构特征。Struc2vec 和 SDNE 两种算法则表现出了相对较低的性能,尽管它们尝试通过结构角色比较或结合自动编码器与图正则化来保留节点的局部和全局结构特性。例如,在 Twitter-Foursquare 数据集上,Struc2vec 的 F1 得分仅为 0.6081,而在 Flickr-Lastfm 数据集上,SDNE 的 F1 得分也只达到了 0.6040。这一结果表明,虽然这两种方法在某些情况下有效,但它们对复杂的社会网络结构可能不够敏感。因此,选择 w-LINE 算法参与后续实验。

5.6 消融实验

为全面评估本文算法中 w-LINE、增强 GNNs、ADCN 这 3 个创新部分对整体性能的具体贡献,开展了一系列消融实验。实验不仅包括单独移除每个组成部分来观察其影响,还会探索不同组合下各部分之间的相互作用,展示它们如何通过协同以提升算法效能。实验结果如表 4、表 5、图 8、图 9 所示。

表 4 Twitter-Foursquare 上的消融实验结果

Table 4 Ablation experimental results of Twitter-Foursquare						
实验组	节点嵌入算法	GNNs	CN	P	R	F1
A1	w-LINE	I 类	N1	0.9274	0.9011	0.9141
A2	w-LINE	II 类	N1	0.8365	0.8319	0.8342
A3	w-LINE	I 类	N2	0.8163	0.8245	0.8204
A4	w-LINE	II 类	N2	0.7047	0.7289	0.7166
A5	LINE	I 类	N1	0.7142	0.7264	0.7202
A6	LINE	II 类	N1	0.6527	0.6418	0.6472
A7	LINE	I 类	N2	0.6619	0.6322	0.6467
A8	LINE	II 类	N2	0.5912	0.6190	0.6048

表 5 Flickr-Lastfm 上的消融实验结果

Table 5 Ablation experimental results of Flickr-Lastfm						
实验组	节点嵌入算法	GNNs	CN	P	R	F1
B1	w-LINE	I 类	N1	0.8137	0.8228	0.8182
B2	w-LINE	II 类	N1	0.7284	0.7332	0.7308
B3	w-LINE	I 类	N2	0.7256	0.7018	0.7135
B4	w-LINE	II 类	N2	0.6362	0.6239	0.6230
B5	LINE	I 类	N1	0.6513	0.6457	0.6485
B6	LINE	II 类	N1	0.5971	0.6475	0.6213
B7	LINE	I 类	N2	0.5852	0.5834	0.5843
B8	LINE	II 类	N2	0.5024	0.5203	0.5112

注:“I 类”表示增强 GNNs,“II 类”表示普通 GNNs;“N1”表示 ADCN,“N2”表示普通 CN。

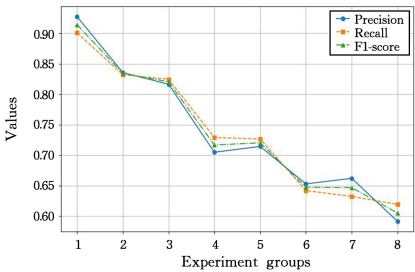


图 8 Twitter-Foursquare 上的消融实验结果
Fig. 8 Ablation experimental results of Twitter-Foursquare

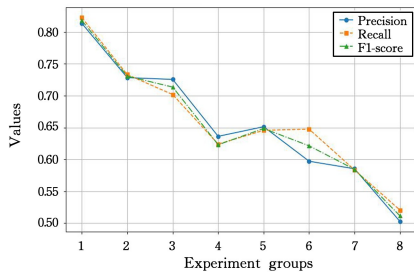


图9 Flickr-Lastfm上的消融实验结果

Fig.9 Ablation experimental results of Flickr-Lastfm

由以上结果可知,在 Twitter-Foursquare 数据集上,当采用 w-LINE 作为节点嵌入算法,并结合增强型 GNNs 和 ADCN 时,模型在 P,R 及 F1 分数方面均取得了最优表现,其中 F1 值达到了 0.9141;而在 Flickr-Lastfm 数据集上,相同配置也实现了最佳的 F1 得分。

通过对比移除或替换各组件后的结果,可以观察到每一个组件对整体性能都有显著影响。例如,将 w-LINE 替换为传统的 LINE 算法后,无论是与增强型还是普通类型的 GNNs 配合使用,或者是在 ADCN 或 CN 下运行,所有测试指标均有明显下降。这表明,w-LINE 在提高节点表示质量方面具有不可替代的作用。同样地,增强型 GNNs 相较于普通 GNNs,在保持其他条件不变的情况下,能够带来更优的预测结果,证明其在信息传播和聚合机制上的改进是有效的。此外,ADCN 的应用也体现出其相对于普通 CN 的优势,特别是在捕捉复杂网络结构中的特性时,ADCN 进一步增强了模型的表达能力和泛化能力。w-LINE 与增强 GNNs 的结合也体现出强大的协同效应。w-LINE 生成的高质量节点嵌入为增强 GNNs 提供了更丰富的特征输入,有助于后者更有效地学习节点间的关系。这种组合不仅提高了模型对节点属性的学习效率,而且在处理多跳邻居信息时也表现出了较强的能力,从而提升了模型的整体预测精度。综上所述,各个组件在单独及协同工作时,对提升算法效能具有关键作用。

结束语 本文提出的基于隐藏特征的增强 GNNs 跨网络身份关联算法,在不同数据集的各项实验指标中都取得了较好的结果,但仍然存在不足,有待进一步优化。首先,w-LINE 虽然算法在捕捉节点间的共现概率方面表现优异,但主要依赖于度较大的节点和一阶邻居节点的信息,这意味着对于那些度较小或处于社交网络边缘的节点,该算法可能无法提供同样精确的身份关联结果。其次,增强 GNNs 的设计虽然增强了模型处理复杂网络结构的能力,但在处理极度稀疏或高度动态变化的网络时,仍可能面临表达能力不足的问题。最后,自适应胶囊网络虽然能够灵活应对不同的正负样本分布情况,但对于某些极端不平衡的数据集,其性能可能会受到一定影响。

下一步,将综合利用多模态信息,更全面地捕捉用户的行为模式和特征,显著提升身份关联的准确性和鲁棒性。结合先进的图神经网络技术与跨模态学习策略,有望突破传统方法在处理单一类型数据时面临的局限,进一步提升跨社交网络身份关联的能力。

参考文献

- [1] WEI W,HUANG C,XIA L,et al. Contrastive meta learning with behavior multiplicity for recommendation[C] // Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022:1120-1128.
- [2] LI N,TSIGKANOS C,JIN Z,et al. Early validation of cyber-physical space systems via multi-concerns integration[J]. Journal of Systems and Software,2020,170:110742.
- [3] GAN Y,ZHANG C F,YANG R S. User identity alignment across heterogeneous networks based on meta-path attention [C]// International Conference on Computer Application and Information Security(ICCAIS 2021). 2022.
- [4] LI Y,LIU Q. A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments[J]. Energy Reports,2021,7:8176-8186.
- [5] CHEN L,CHEN J,XIA C. Social network behavior and public opinion manipulation[J]. Journal of Information Security and Applications,2022,64:103060.
- [6] LIU J L,LIU Y,MA C X,et al. User Identity Parsing in Heterogeneous Social Platforms [J]. Data Collection and Processing, 2022,37(5):1101-1114.
- [7] ZHANG J,GUO Y G. Social Network User Identification Method based on Preference Logic [J]. Computer Simulation, 2022,39(4):450-453,505.
- [8] LIU Z L,QIN T,GUAN X H,et al. Online user identity attribute association method using user name similarity propagation model [J]. Journal of Xi'an Jiaotong University,2016,50(4):1-6,27.
- [9] ANISA H,AYDAY E. Profile matching across online social networks[C]// Information and Communications Security; 22nd International Conference. Springer, 2020.
- [10] LI Y,PENG Y,JI W,et al. User identification based on display names across online social networks[J]. IEEE Access,2017,5: 17342-17353.
- [11] DING X,ZHANG H,MA C,et al. User identification across multiple social networks based on naive Bayes model[J]. IEEE Transactions on Neural Networks and Learning Systems,2024, 35:4274-4285.
- [12] QU Y T,XING L,MA H Z,et al. Exploiting user friendship networks for user identification across social networks[J]. Symmetry,2022,14(1):110.
- [13] LI Y J,PENG Y,JI W L,et al. User identification based on display names across online social networks [J]. IEEE Access, 2017,5:17342-17353.
- [14] DAI J,MA Q. Cross-social network user matching based on user Checkin [J]. Computer Engineering and Applications, 2023, 59(2):76-84.
- [15] HUO T F. Cross-social media user identity linking based on deep modeling of user behavior [D]. Beijing: University of Chinese Academy of Sciences,2022.
- [16] LI Y J,JI W L,GAO X,et al. Matching user accounts with spatio-temporal awareness across social networks[J]. Information Sciences,2021,570:1-15.

[17] ZHENG C H,LI P,PENG W,JORA:Weakly supervised user identity linkage via jointly learning to represent and align[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024,35(3):3900-3911.

[18] SENETTE C,MARCO S,MAURIZIO T. User Identity Linkage on Social Networks;A Review of Modern Techniques and Applications[J]. IEEE Access,2024,12:171241-171268.

[19] DING F X,MA X Q,YANG Y,et al. User identity linkage across location-based social networks with spatio-temporal check-in patterns[C]// 2020 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/ SustainCom). IEEE,2020.

[20] HANG Z B,GU Q H,YUE T,et al. Identifying the same person across two similar social networks in a unified way;Globally and locally[J]. Information Sciences,2017,394(C):53-67.

[21] KONG X,ZHANG J,YU P S. Inferring Anchor Links across Multiple Heterogeneous Social Networks[C]// The 22nd ACM International Conference on Information & Knowledge Management,2013;179-188.

[22] MA X,DING F,PENG K,et al. CP-link;Exploiting continuous spatio-temporal check-in patterns for user identity linkag[J]. IEEE Transactions on Mobile Computing, 2023, 22(8):4594-4606.

[23] LIU L,CHEN P,LI X,et al. Wlalign;Weisfeiler-lehman relabeling for aligning users across networks via regularized representation learning[J]. IEEE Transactions on Knowledge and Data Engineering,2023,36(1);445-458.

[24] CHEN H,YIN H,SUN X,et al. Multi-level Graph Convolutional Networks for Cross-platform Anchor LinkPrediction[C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2020;1503-1511.

[25] CHU X,FAN X,YAO D,et al. Cross-network embedding for multi-network alignment[C]// The World Wide Web Conference. 2019;273-284.

[26] MAN T,SHEN H,LIU S,et al. Predict anchor links across social networks via an embedding approach[C]// IJCAI. 2016: 1823-1829.

[27] KELKHA M M,RAHGOZAR M,ASADPOUR M. DeepLink: A novel link prediction framework based on deep learning[J]. Journal of Information Science 2021,47(5):642-657.

[28] CHEN W,WANG W Q,YIN H Z,et al. HFUL;a hybrid framework for user account linkage across location-aware social networks[J]. The VLDB Journal,2023,32(1);1-22.

[29] ZHOU F,WEN Z J,ZHONG T,et al. Unsupervised user identity linkage via graph neural networks[C]// GLOBECOM 2020-2020 IEEE Global Communications Conference. IEEE,2020.

[30] XIONG X,XIE X,WU Y,et al. DSANE;A dual structure-aware network embedding approach for user identity linkage[C]// Proceedings of the IEEE 8th International Conference on Big Data Analytics(ICBDA). 2023;193-198.

[31] LI S D,LU D N,LI Q,et al. MFLink; User identity linkage across online social networks via multimodal fusion and adversarial learning[J]. IEEE Transactions on Emerging Topics in Computational Intelligence,2024,8(5):3716-3725.

[32] HUANG S,XIANG H,LENG C,et al. Cross-Social-Network User Identification Based on Bidirectional GCN and MNF-UI Models[J]. Electronics,2024,13;2351.

[33] ZHOU F,LIU L,ZHANG K P,et al. DeepLink;A Deep Learning Approach for User Identity Linkage[C]// IEEE Conference on Computer Communications. 2018;1313-1321.

[34] TANG W,SUN H,WANG J,et al. Identifying users across social media networks for interpretable fine-grained neighborhood matching by adaptive GAT[J]. IEEE Transactionas on Services Computing,2023,16(5):3453-3466.

[35] FENG J,ZHANG M Y,WANG H D,et al. DPLink; User identity linkage via deep neural network from heterogeneous mobility data[C]// The Web Conference. 2019;459-469.

[36] ZHANG J,CHEN B,WANG X M,et al. MEgo2Vec; Embedding matched ego networks foruser alignment across social networks. [C]// International Conference on 16 Journal of Frontiers of Computer Science and Technology Information and Knowledge Management. 2018;327-336.

[37] SHAO J,WANG Y,GAO H,et al. AsyLink;user identity linkage from text to geo-location via sparse labeled data[J]. Neurocomputing. 2023,515;174-184.

[38] ZHANG J W,YU S P. PCT;Partial co-alignment of social networks[C]// The Web Conference. 2016;749-759.

[39] ZHANG Y T,TANG J,YANG Z L,et al. COSNET;Connecting heterogeneous social networks with local andglobal consistency [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2015;1485-1494.

PAN Yuquan, born in 2001, postgraduate. His main research interest is social network analysis.

YUAN Deyu, born in 1986, Ph.D, associate professor, master’s supervisor. His main research interests include network security and social network analysis.

(责任编辑:何杨)