

基于同质多层图卷积的多尺度网络对齐模型

崔家豪, 江涛, 徐梦瑶

(西北民族大学语言与文化计算教育部重点实验室, 甘肃 兰州 730030)

摘要: 社交网络对齐作为网络科学的重要研究方法, 已在多个领域广泛应用。现有方法通常依赖高质量的用户属性信息来完成特定任务, 但隐私保护机制的存在使得这些信息难以获取。此外, 若仅依赖网络拓扑结构, 可能面临数据不足的挑战。针对上述问题, 基于节点邻域特征和网络同质性提出一种基于同质多层图卷积的多尺度网络对齐模型。节点特征方面, 通过K近邻算法聚合节点邻域信息建模深层网络结构, 从而进行数据增强。图卷积方面, 根据网络同质性构建同质度矩阵, 对卷积过程进行引导, 并以网络社区结构为基础来处理不同尺度的社交网络。在两个不同规模的现实社交网络实验结果表明, 该方法能够有效提升社交网络对齐任务的性能。

关键词: 网络对齐; 网络嵌入; 网络同质性; 多尺度网络; 社区发现

中图分类号: TP391.1

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202445

Multiscale network alignment model based on convolution of homogeneous multilayer graphs

CUI Jiahao, JIANG Tao, XU Mengyao

Key Laboratory of Language and Cultural Computing of Ministry of Education, Northwest Minzu University, Lanzhou 730030, China

Abstract: Social network alignment as an important research method in network science has been widely used in several fields. Existing methods usually rely on high-quality user attribute information to complete specific tasks, but the existence of privacy protection mechanisms makes this information difficult to obtain. In addition, relying solely on network topologies can be challenged by insufficient data. In order to solve the above problems, a cross-network user alignment model based on the node neighborhood characteristics and network homogeneity was proposed. In terms of node characteristics, the K-nearest neighbor algorithm was used to aggregate node neighborhood information to model the deep network structure, so as to enhance the data. In terms of graph convolution, the convolution process was guided by the construction of a homogeneity matrix according to the network homogeneity, and the social networks of different scales were processed based on the network community structure. Experimental results on two real-world social networks of different scales show that the proposed method can effectively improve the performance of social network alignment tasks.

Key words: network alignment, network embedding, network homogeneity, multi-scale network, community discovery

收稿日期: 2024-08-19; 修回日期: 2024-11-04

通信作者: 江涛, xinxiyuanjt@126.com

基金项目: 中央高校基本科研业务费项目 (No.31920240081, No.31920240127)

Foundation Items: Fundamental Research Funds for the Central Universities of Northwest Minzu University (No.31920240081, No.31920240127)

0 引言

在线社交网络是当今互联网最活跃的组成部分, 不仅仅是人们沟通交流的平台, 还是信息传播、影响力扩散和社会联系的重要媒介。随着社交网络的迅速发展和普及, 人们在不同社交网络上创建的个人资料、社交关系和交互行为变得越来越复杂。社交网络平台的多样性以及用户行为的复杂性, 导致同一用户在不同社交网络之间的个人资料和社交关系存在着差异, 但在不同社交平台上仍然存在着大量重叠的个人用户账户。受到用户个人隐私限制, 大多数社交网络中并未明确给出属于同一个人的多个账户的信息, 导致不同网络中的相同用户彼此间的关联不可知。

社交网络对齐作为解决用户同一身份认定问题的一种关键技术, 旨在将不同社交网络中相似的用户通过其社交关系对齐到一起, 从而实现跨社交网络的数据整合与分析。这种对齐可以帮助研究人员理解用户在不同社交网络上的行为和关系, 有助于改进在线个性化服务^[1]、跨网络信息传播^[2]、链接预测^[3]、推荐系统^[4-5], 以及犯罪行为检测^[6]等下游网络分析任务。尽管人们对这个具有挑战性课题的关注不断提高并提出众多解决方案, 但仍存在很大的改进空间。传统的研究方法^[7-10]通过利用用户可见的辅助信息来解决这个问题, 例如用户自我定义的个人资料、每日发布的内容以及其他人口统计特征(用户名、个人资料图片、位置、性别、帖子、博客、评论等)。然而, 随着公众隐私和信息权意识的增强, 这些信息变得越来越难以获取。

随着网络嵌入技术的发展, 与此问题相关的研究重点已逐渐转移到如何利用该技术有效挖掘网络结构信息。与可见的辅助信息相比, 社交网络结构数据在数据正确性和完整性方面更加可靠。鉴于此, 本文提出了一种基于同质多层图卷积的多尺度网络对齐模型, 该模型包括同质多层图卷积网络(homogeneous multilayer graph convolutional network, HMGCN)以及基于该网络构建的多尺度网络对齐框架。现有方法多关注网络结构本身的建模方法, 因为对于不同尺度社交网络来说, 网络结构信息与网络内部节点特征信息也不尽相同。网络过于稀疏会导致网络结构信息不足, 本文提出的HMGCN通过融合单一社区网络内结构表层及深层特征解决了多尺度网络对齐中存在的限制问题。具体来说,

HMGCN利用K近邻(K-nearest neighbor, KNN)算法从原始网络结构中挖掘节点邻域信息进行数据增强, 再根据节点特征构建同质度矩阵来引导图卷积过程, 优化网络邻域内节点特征传播和聚合。除此之外, 随着网络规模的增加, 网络嵌入技术也面临GPU计算内存不足的问题。因此本文提出了一个多尺度网络对齐框架来实现网络对齐。网络社区是解决这一问题的关键, 在每个社区中分别部署本文的模型来处理不同尺度的社交网络。本文的主要贡献如下。

(1) 提出一种能够有效融合网络结构与节点特征等信息的模型, 聚合网络节点邻域信息并使用同质度矩阵引导网络卷积, 优化网络邻域内节点特征传播。

(2) 从网络社区角度入手, 将社区结构视为单一网络, 将不同社区转化为低维向量表示, 其后以向量空间联合机制实现网络对齐, 两者作为联合框架解决网络对齐中大规模社交网络计算方面存在的限制。

(3) 对不同尺度社交网络数据集进行实验, 结果表明, HMGCN可以很好地从节点特征和拓扑结构中提取足够的相关信息, 以满足多尺度网络对齐的需求。

1 相关工作

1.1 基于用户属性信息的网络对齐方法

在社交网络的用户对齐研究中, 传统方法往往依赖于用户的属性信息, 如用户名、教育经历和居住城市等。然而, 这些属性信息的获取存在隐私问题, 限制了方法的有效性。为了应对这一挑战, 研究者们逐渐转向基于用户轨迹数据和网络结构信息的算法。Riederer等^[10]利用社交网络中丰富的用户位置信息提出了一种自适应算法, 该算法从用户的轨迹数据出发, 分析用户对的相似性, 用于实现跨网络用户对齐。在社交网络对齐任务研究初期, 一些研究人员提出了大量基于用户画像信息^[11-12]的跨社交网络用户对齐方法。用户画像信息包括有关两个网络中用户的各种个人资料信息(用户名、教育经历、居住城市、个人描述等)。结果表明, 基于用户档案信息的跨社交网络用户对齐模型对于某些社交网络是可行且有效的。Zhang等^[13]基于代价敏感性提出了一种算法, 通过提取一些用户特征(前缀、后缀、用户名稀有程度、用户习惯等)来衡量

不同社交网络之间的用户相似度。Zhao等^[14]提出了一种用于社交网络对齐的反向传播神经网络(back propagation neural network, BPNN), 利用反向传播神经网络获取跨社交网络的用户名向量之间的映射, 将分类问题转变为向量之间的映射问题, 提高了社交网络对齐任务的准确率。然而, 在实际的社交网络场景中, 用户的属性信息很难获取, 这涉及用户的隐私。许多用户的属性信息无法访问, 并且某些用户会出于各种目的去模仿他人或伪造个人信息, 使得基于用户属性信息的用户对齐方法难以发挥作用。本文提出的HMGCN基于网络结构信息并结合网络结构表面及深层特征优化节点表示, 而且完全不依赖用户属性信息, 从而克服了传统方法的局限性。

1.2 基于网络结构的网络对齐方法

现有的使用网络结构进行社交网络对齐的研究工作可以分为两大类。一类直接关注社交网络结构的表面特征(度、聚类系数、共同邻居等)^[15-18], 通过衡量节点之间的相似性, 帮助模型更好地学习节点表示。另一类通过图嵌入方法^[19-21]获得网络结构的深层特征, 并根据深层特征进行社交网络对齐, 该类方法可以更好地保留网络的内在规律性。为了获得各个网络节点的深层特征, 这些方法首先在不同网络上采用图嵌入。接下来, 通过可观察的锚链接将跨网络节点对的深层特征约束到公共潜在特征空间中。最后, 使用公共潜在特征空间来识别跨网络节点对是否为相同用户。

1.2.1 关注社交网络结构表面特征的方法

在社交网络对齐的研究中, 提取社会特征是一种有效的策略。Zhan等^[15]提出从网络中提取各种社会特征, 包括扩展的共同邻居、扩展的Jaccard系数和Adamic/Adar指数。这些特征考虑了共享邻居的数量、用户邻域的大小以及不同公共邻居的权重。此外, 通过捕获有关用户位置、时间活动以及他们在帖子中使用的单词的信息来提取空间分布、时间分布和文本使用等特征, 以此来实现网络对齐。Zafarani等^[17]提出了一种通过分析用户名在不同社交媒体网站上的使用情况来连接用户的方法。网络结构表面特征往往反映用户的个人爱好以及行为偏好, 能有效缓解网络对齐领域面临的数据不足问题。尽管提取社交网络的结构表面特征在用户对齐中发挥了重要作用, 但这一方法也存在一些不足。首先, 这些特征主要依赖于共享邻居和用户行

为的表面信息, 可能无法捕捉用户之间更深层次的关系和互动模式。其次, 社交网络的动态性和用户行为的多样性使得仅依靠表面特征难以适应不断变化的网络环境。

1.2.2 关注社交网络结构深层特征的方法

在结合网络结构深层特征的网络对齐研究中, 如何有效提取和整合网络信息以增强模型性能成为关键。Chen等^[21]通过整合从原始网络中提取的超图信息进行数据增强, 并提出了一种新颖的嵌入框架多级图卷积网络, 以联合学习不同粒度级别的网络节点的嵌入。对于两个及两个以上的多网络问题, Chu等^[22]提出的跨多网络对齐(cross multi-network alignment, CrossMNA)模型考虑了多网络场景, 其中涉及至少两种具有不同网络结构的网络, 利用跨网络信息来细化两种类型的节点嵌入向量, 即用于网络对齐的向量间向量和用于其他下游网络分析任务的向量内向量, 以此实现多网络对齐任务。Saxena等^[23]提出一种用于对齐结构异构网络的框架, 该框架使用一组标记(映射)锚节点来学习映射节点的结构属性的相关性。跨网络交换的信息有助于增强图神经网络的表达能力, 从而处理任何潜在的过度平滑问题。Li等^[24]提出了一种基于三层注意机制的网络嵌入(triple-layer attention mechanism-based network embedding, TANE)方法, 利用一阶和二阶邻居的加权结构相似度来学习潜在结构信息, 以降低网络稀疏度, 通过充分挖掘网络结构来识别用户。Lu等^[25]基于图神经网络和双层注意力机制来学习用户节点在社交网络中的嵌入向量, 而后基于嵌入向量提出双向对齐策略, 预测源社交网络与目标社交网络之间的锚链接。本文提出的HMGCN在结合单一社区网络邻接矩阵进行数据增强学习的同时, 挖掘网络内部同质信息以优化表示学习。社交网络深层结构需要结合不同领域相关理论对其蕴含的信息进行深度挖掘, 这也是后续研究中需要关注的。

2 多尺度网络对齐方法

2.1 问题定义

给定一对网络 $G_1=\{V_1, E_1\}$ 和 $G_2=\{V_2, E_2\}$, 以及一组观察到的锚链接 $S_{\text{anchor}}=\{(u, v)|u \in V_1, v \in V_2\}$, 其中 V_i 代表网络 G_i 中的节点集合, E_i 代表网络 G_i 中的边集合, 网络对齐任务的研究目标是对 G_1 和 G_2 中的用户进行同一身份认定。本文的任务定义

如下：给定一对节点 (u, v) ，其中 $u \in V_1, v \in V_2$ ，学习一种二元分类模型，使其能够正确预测给定的节点对是否为跨网络的相同用户。

2.2 模型概述

HMGCN框架如图1所示。首先，将社区网络视为单一网络结构，并通过社交网络拓扑结构对节点间的连接关系进行建模，以捕获网络数据的拓扑结构信息，并生成节点特征矩阵。在拓扑结构建模的基础上，引入K近邻矩阵，通过衡量节点之间的相似性，帮助模型更出色地学习节点表示并扩充节点邻域特征信息，与网络拓扑结构信息进行融合学习。为了更好地利用网络局部结构的优势，本文在模型中引入了同质度矩阵，以引导图卷积操作进行节点表示学习。以此使模型能够适应不同尺度网络，从规模不同以及稠密程度不一的网络中提取足够的网络结构信息以优化表示学习。在模型的前向传播过程中，本文通过定义的损失函数对模型进行优化学习，以最小化模型输出节点表示向量与真实节点表示之间的差异，对网络数据实现最优表示学习。

2.3 网络拓扑图卷积

网络拓扑图卷积是一种用于处理原始社交网络的方法。给定一个社交网络 $G=(V, E)$ ，其中 V 表示

节点集合， E 表示边集合。本文首先对该网络进行社区划分，得到 M 个社区 $C=\{C_1, C_2, \dots, C_M\}$ ，每个社区 C_i 对应一个子图 $G_i=(V_i, E_i)$ ，其中 V_i 是社区 C_i 中的节点集， E_i 是社区 C_i 中的边集， i 表示社区索引。对于每个社区 C_i ，本文将其对应的子图 G_i 的邻接矩阵表示为 $A_i \in \mathbb{R}^{|N_i| \times |N_i|}$ ，其中 N_i 表示社区 C_i 中节点的数量。在图卷积过程中，首先计算支持向量 $\mathbf{adj}_i$ 。

$$\mathbf{adj}_i = X_i W_i \quad (1)$$

其中， X_i 为基于社区 C_i 的网络拓扑结构形成的节点特征矩阵， W_i 为卷积层中共享的可训练权重。将邻接矩阵 A_i 与支持向量 $\mathbf{adj}_i$ 进行矩阵乘积运算，得到更新的消息矩阵 z_i 。

$$z_i = A_i \times \mathbf{adj}_i \quad (2)$$

然后，利用 Sigmoid 激活函数 σ 对每个社区中的消息矩阵 z_i 进行非线性变换，得到更新后的节点特征矩阵 X'_i 。

$$X'_i = \sigma(z_i) \quad (3)$$

除网络邻接矩阵外，通过构建K近邻矩阵对输入模型进行数据增强。对于原始图 G 划分得到的每个社区 C_i ，基于其对应的子图 G_i 来构建K近邻矩阵 K_i 。首先，对每个子图 G_i 构造其对应的K近邻

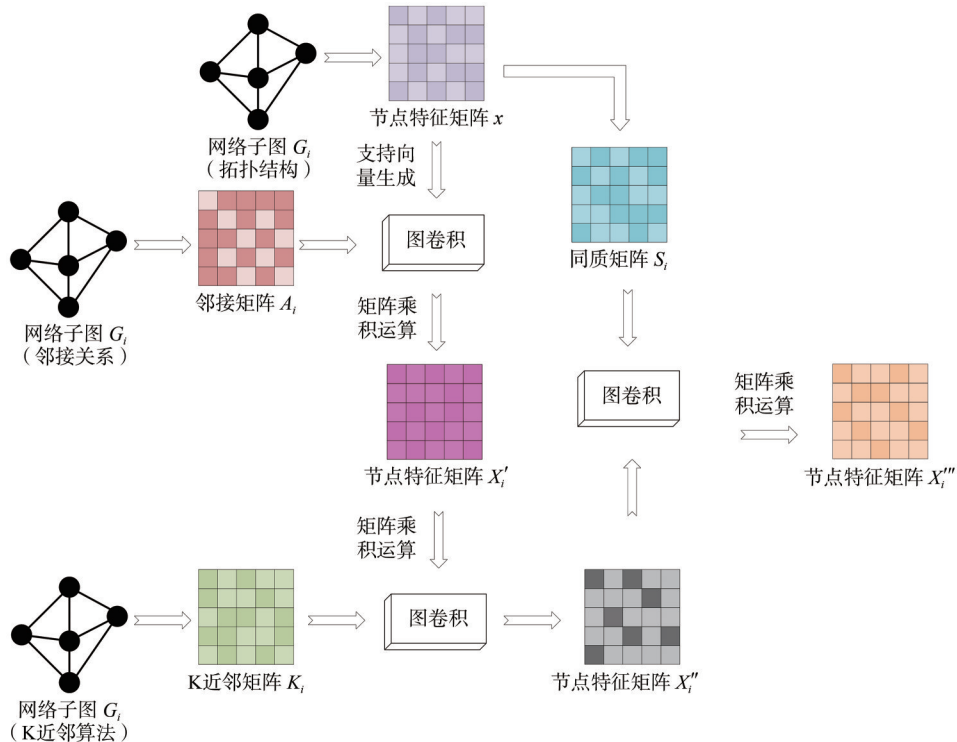


图1 HMGCN框架

图。具体来说,对于每个节点 $v \in V_i$, 找到其 K 个结构相似度最高的邻居节点构成其 K 近邻域, 结构相似度是根据节点与其邻居节点间的结点重合程度来定义的。其中, 本文将每个节点的 K 近邻域结构信息编码为该节点的一系列结构特征, 即对于节点 v , 其第 i 个特征定义为与其第 i 个 K 近邻节点的重合程度。由此构建出子图 G_i 的 K 近邻矩阵 K_i 。对 K_i 进行添加自连接特征和归一化等正则化处理, 从而获得规整的结构特征表达。

本文通过图卷积操作来融合 K 近邻节点特征矩阵, 优化拓扑图卷积的表示, 首先计算支持向量 sadj_i 。

$$\text{sadj}_i = X_i' W_i \quad (4)$$

而后根据 K 近邻矩阵 K_i 传播消息得到更新的消息矩阵 z_i' 。

$$z_i' = K_i \times \text{sadj}_i \quad (5)$$

使用 Sigmoid 激活函数 σ 对每个社区中的消息矩阵 z_i' 进行变换。

$$X_i'' = \sigma(z_i') \quad (6)$$

最后返回更新后的节点特征矩阵 X_i'' , 至此完成对网络拓扑结构以及节点邻域特征的表示学习。

2.4 同质度引导图卷积

同质度不仅可以衡量网络内部节点相似属性, 还可以引导节点邻域特征传播。针对每个社区 C_i 对应的子图 $G_i = (V_i, E_i)$, 定义同质矩阵共享训练权重 Q , Q 为一个 $d \times d'$ 的权重矩阵, 其中 d 是输入节点特征的维度, d' 是隐层表示的维度, 并将 Q 用于节点特征的线性变换。首先, 将输入节点特征矩阵

X_i 与 Q 进行矩阵乘积运算, 得到节点的隐层表示 h 。

$$h = X_i Q \quad (7)$$

然后, 对隐层表示 h 进行 softmax 归一化操作, 并将其转置矩阵与自身相乘, 得到同质矩阵 S_i 。

$$S_i = \text{softmax}(h) \times \text{softmax}(h)^T \quad (8)$$

在引入同质矩阵后, 同质度图卷积操作变为:

$$H = S_i X_i Q \quad (9)$$

最终的节点特征更新为:

$$X_i''' = H + X_i'' \quad (10)$$

其中, 同质矩阵 S_i 捕获了节点之间的同质关系视角, X_i'' 保留了节点之间的直接连接视角与邻域连接视角, 将同质矩阵 S_i 和 X_i'' 进行加和, 以融合多个视角的信息, 从而进行引导结构信息传播。

2.5 多尺度网络对齐框架

在线社交网络平台中多尺度的社交网络在规模、密度以及内部连接上都存在巨大差异, 为使模型能够处理不同尺度的社交网络, 本文基于社区发现算法将网络划分为大小相似的社区, 在此基础上提取不同网络社区的结构矩阵与特征矩阵, 并通过构建同质度矩阵引导模型进行网络拓扑结构与节点特征融合。为了对齐不同社区学习嵌入空间, 通过社区内共享节点将各社区嵌入向量映射至同一向量空间内, 最后将源网络的向量空间通过线性层映射至目标网络向量空间内, 对齐两个不同的网络, 以实现更加准确的多尺度网络对齐。多尺度网络对齐框架如图2所示。

2.5.1 社区发现算法

Louvain 算法^[26]是一种基于模块度的社区发现

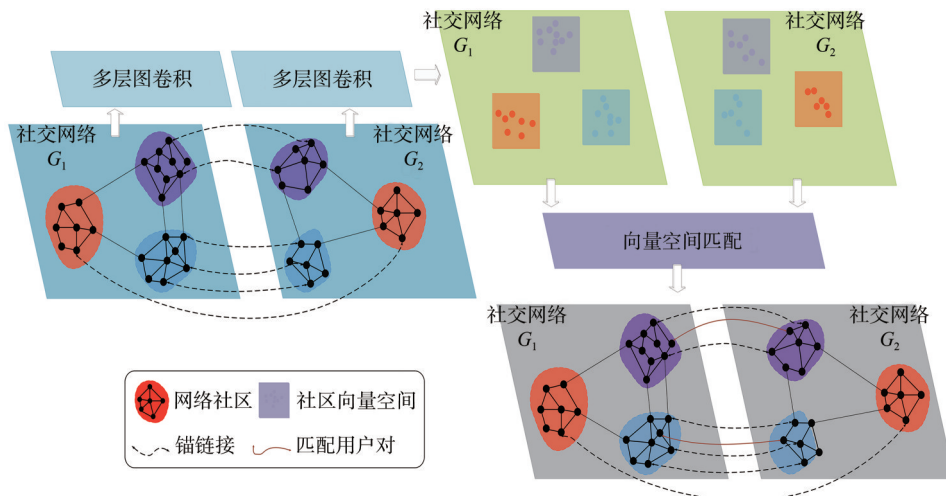


图2 多尺度网络对齐框架

算法。其基本思想是网络中节点尝试遍历所有邻居的社区标签，并选择最大化模块度增量的社区标签。在最大化模块度之后，将每个社区看成一个新的节点，重复Louvain算法直到模块度不再增大。

为了将大型网络分割成可接受大小的若干分区（例如，从 $N_{\min}$ 节点到 $N_{\max}$ 节点），本文首先利用Louvain算法来计算最大化模块化的网络分区。对于每个子图 $G_i=(V_i, E_i)$ ，将其视为初始分区。如果大小大于上界，即 $|V_i|>N_{\max}$ ，需要再次将 G_i 作为输入，并重复该算法，将 G_i 进一步拆分成更小的分区。如果 $|V_i|<N_{\min}$ ，则将其随机分配给其他创建的分区，直到网络内各分区规模合理。

2.5.2 社区向量空间联合

由于不同社区嵌入表示是在不同向量空间中，所以在获得两个网络 G_1 和 G_2 的不同社区嵌入表示之后，不应该直接将它们用于社交网络对齐任务，从网络全局结构角度看，只对部分社区进行对齐的方法是欠缺考虑的。为此从网络中选择 N 个节点作为所有分区的共享节点，并将这些 N 节点及其关联的边附加到所有分区中。然后选择其中一个分区作为固定分区，并将其他分区协调到与其相同的空间中，其具体过程如图3所示。具体来说，对于社区集合 $C=\{C_1, C_2, \dots, C_k\}$ ，通过固定社区 C_1 ，所有其他划分的嵌入通过线性映射函数进行变换。本文最大限度地实现以下向量联合匹配目标。

$$O_{\text{partition}} = \sum_{p=2}^P \sum_{v_i \in \mathcal{V}_{\text{shared}}} \log \sigma \left((f_p(x_i^{(p)}))^T x_i^{(1)} \right) \quad (11)$$

其中， $\mathcal{V}_{\text{shared}}$ 是出现在所有分区中的共享节点的集合， $x_i^{(p)}$ 是节点 v_i 在分区 p 中的嵌入表示， P 为单个网络中的分区数量， $f_p(\cdot)$ 是转换函数，使得该分区中的节点嵌入经过转换后与第一个分区中的相应节点嵌入尽可能相似。将每个社区匹配到相同的空间中，就可以得到单一网络的联合网络嵌入。

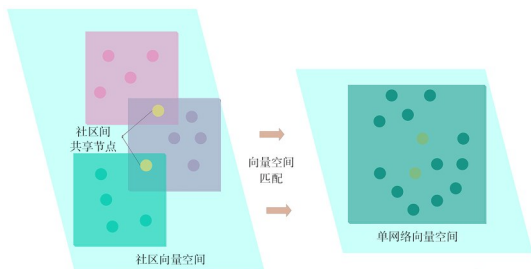


图3 社区向量空间联合过程

2.5.3 网络向量空间匹配

为了将不同网络嵌入表示匹配至同一向量空间，以便于下游网络对齐任务的进行，本文通过两个线性全连接层对输入的不同网络嵌入向量进行变换。固定 X_1^K 嵌入向量，并将 X_2^K 投影到与 X_1^K 相同的空间中，从而达到将 X_1^K 和 X_2^K 协调到同一空间中的目的，其中 X_1^K 和 X_2^K 表示不同网络的嵌入向量。设 $\phi(\cdot|\Gamma, \mathbf{b})$ 是一个投影函数，投影矩阵为 Γ ，偏置为 $\mathbf{b}$ 。然后，通过对齐两个图中锚节点的嵌入向量，本文可以学习投影函数中的参数，从而确保两个潜在空间的准确协调。

$$O_{\text{anchor}} = \sum_{(v, u) \in S_{\text{anchor}}} \left\| X_1^K[v, :] - \phi(X_2^K[u, :]|\Theta, \mathbf{b}) \right\|^2 \quad (12)$$

其中， S_{anchor} 是标记的锚链接， O_{anchor} 为预测目标锚链接， Θ 为模型参数。对于任何节点对 $\{(v_i, v_j)|v_i \in G_1, v_j \in G_2\}$ ，其可以通过它们对应的嵌入的级联来表示。将这些节点对嵌入一个全连接网络中，通过全连接网络的输出判断它们是否为锚链接的预测，并使用交叉熵函数作为锚链接预测的损失函数来优化模型。

2.6 多尺度网络对齐框架复杂度分析

在设计和优化大规模网络处理框架时，复杂度分析是确保模型效率和可扩展性的关键步骤。本文所提出的HMGCN及多尺度网络对齐框架，旨在通过社区化结构来优化计算和内存管理。为了深入理解该框架的性能，本节对其时间复杂度和空间复杂度进行了详细分析。具体而言包括以下3个方面：总体计算复杂度如何受到网络划分的影响、社区化如何通过并行处理提升计算效率，以及如何通过分散内存需求来优化空间使用。

对于时间复杂度而言，当网络为整体化结构时，整体框架的时间复杂度为 $O(NdL)$ ，其中， N 是网络中的节点总数， d 是每个节点的特征维度， L 是图卷积的层数。在未进行社区化的情况下，这一复杂度反映了整个网络处理的计算量。而当网络被划分为 C 个社区时，每个社区的规模为 $\frac{N}{C}$ ，每个社区内部的计算复杂度为 $LO\left(\frac{N}{C}d\right)$ 。在并行处理情况下，总体计算时间可缩短为 $\frac{LO(Nd)}{C}$ ，这表明社区化通过并行处理显著提升了计算效率。

在空间复杂度上对该框架进行分析，网络内存

需求为 $O(Nd)$ ，这是因为需要存储整个网络中所有节点的特征信息。不同规模网络社区化后，每个子社区的空间需求为 $O\left(\frac{N}{C}d\right)$ ，社区化使得内存需求被分散到多个子社区中，从而优化了内存管理和实际使用效率。

3 实验设计与结果分析

3.1 实验数据集

本文采用社交网络对齐任务中公开的两个跨平台数据集：学术网络 ArnetMiner-LinkedIn^[27]数据集和社交网络 Weibo-Douban^[28]数据集。ArnetMiner-LinkedIn 数据集包含 4 038 602 个节点，其中 ArnetMiner 网络包含 1 053 188 个节点和 3 916 907 个社交链接，而 LinkedIn 网络包含 2 985 414 个节点，包含 25 965 384 个社交链接连接。在 ArnetMiner-LinkedIn 数据集中，跨两个网络识别了 4 269 个对齐的用户对。Weibo-Douban 数据集连接了中国两个流行的社交媒体平台，即 Weibo（包含 9 714 个节点和 117 218 个社交链接）和 Douban（包含 9 526 个节点和 120 245 个社交链接）。在 Weibo-Douban 数据集中，这两个网络的总共 19 240 个节点中有 1 397 个对齐用户。

根据相关工作^[8-19]，本文将锚链接预测视为二元分类任务。具体来说，以不同网络中的一组节点对 (u, v) 作为输入，本文的目标是预测它们是否代表两个网络中的同一实体。

3.2 评价指标

本文将宏观精度（Marco Precision）、宏观召回率（Marco Recall）和宏观 F1（Macro F1）分数作为实验结果评价指标。以下为各评价指标的计算方式。

宏观精度是指每个类别精度的平均值，即它是对各个类别精度的无权平均（不考虑类别的样本数）。对每个类别，计算其精度（Precision）。

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (13)$$

其中， TP_i 是类别 i 的真正例数， FP_i 是类别 i 的假正例数。

所有类别精度的平均值为：

$$\text{Macro Precision} = \frac{1}{C} \sum_{i=1}^C \text{Precision}_i \quad (14)$$

其中， C 是类别数。

宏观召回率是指每个类别的召回率的平均值。类似于宏观精度，它是各类别召回率的无权平均。对每个类别，计算其召回率（Recall）。

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (15)$$

其中， TP_i 是类别 i 的真正例数， FN_i 是类别 i 的假负例数。

所有类别召回率的平均值为：

$$\text{Macro Recall} = \frac{1}{C} \sum_{i=1}^C \text{Recall}_i \quad (16)$$

宏观 F1 分数是指每个类别的 F1 分数的平均值。F1 分数是精度和召回率的调和平均数，衡量了模型的综合性能。对每个类别，计算其 F1 分数：

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (17)$$

所有类别 F1 分数的平均值为：

$$\text{Macro F1} = \frac{1}{C} \sum_{i=1}^C \text{F1}_i \quad (18)$$

3.3 实验参数设置

对于锚链接预测，训练和测试时正锚链接和负锚链接的比例均设置为 1:1。本文使用 50% 的正向和负向链接来训练所有方法，并在剩余部分对其进行测试。在社区发现过程中，对于 ArnetMiner-LinkedIn 数据集，本文设置 $N_{\min}=1\ 000$ 、 $N_{\max}=15\ 000$ ；对于 Weibo-Douban 数据集，本文设置 $N_{\min}=100$ 、 $N_{\max}=1\ 500$ 。在构建 K 近邻矩阵的过程中，本文设置 $K=5$ 。在模型进行图表示学习的过程中，学习率和嵌入维度分别固定为 0.0001 和 200。

3.4 基线模型

多级图卷积网络（multi-level graph convolutional networks, MGCN）^[21]：MGCN 采用局部图卷积和超图卷积来学习图嵌入，可以捕获更广泛、更丰富的图结构特征。

基于嵌入锚链接预测模型（predict anchor links via embedding model, PALE）^[19]：PALE 通过最大化观察到的边缘和潜在空间匹配的对数似然以及网络嵌入来预测锚链接。

图卷积网络（graph convolutional networks, GCN）^[29]：GCN 是一种能够直接作用于图并且利用其结构信息的卷积神经网络。

CrossMNA^[22]：该模型使用网络间向量来区分多网络间整体结构特征，利用网络内向量来优化细粒度多网络对齐。

3.5 对比实验分析

为验证本文提出的HMGCN的有效性, 本节在ArnetMiner-LinkedIn数据集和Weibo-Douban数据集上将HMGCN与几种基线模型进行对比实验, 实验结果如图4和图5所示。

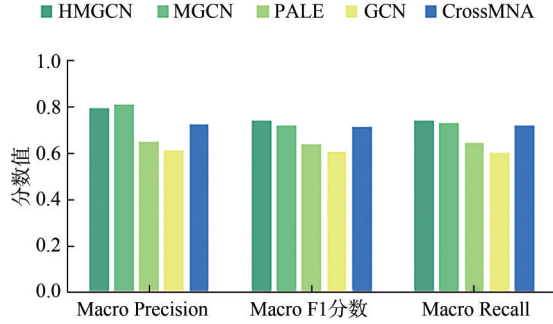


图4 ArnetMiner-LinkedIn数据集对比实验结果

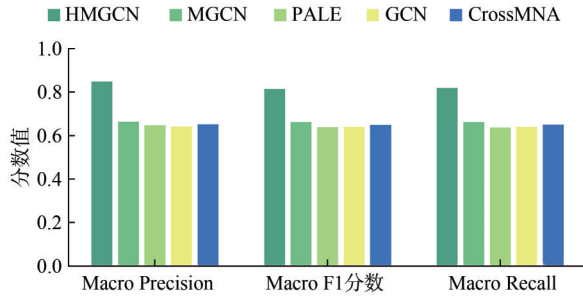


图5 Weibo-Douban数据集对比实验结果

如上文所述, 现有的网络对齐方法在面对不同尺度的数据集时鲁棒性较差, 模型性能不够稳定。而本文所提出的HMGCN通过探索网络结构同质性和提取的K近邻图的结构信息来提高模型在面对不同尺度数据集时的有效性。由图4和图5可知, 就所有评估指标而言, 本文模型在不同尺度的社交网络都展示出了优越的性能。

在ArnetMiner-LinkedIn数据集即大规模社交网络数据集上, 相比于MGCN, 本文提出的HMGCN的Macro Precision略低2.01%, 而在Macro Recall和Macro F1分数上分别提高了1.03%和1.82%。MGCN通过执行局部图卷积和超图卷积操作来实现较好的大规模网络对齐效果, 而HMGCN则是利用节点邻域特征对网络结构信息进行增强, 再利用社区内部同质性来引导图卷积过程, 从而实现高效大规模网络对齐。CrossMNA使用多网络间向量实现多网络对齐, 并以此突破网络对齐之间成对学习的限制, 但HMGCN在网络内表示学习效果明显优于CrossMNA。相比于CrossMNA, HMGCN在Macro

Precision、Macro Recall和Macro F1分数上分别提升了6.81%、2.35%和2.2%, 这3种基线模型相较于HMGCN无法有效处理大规模社交网络复杂的网络拓扑结构。相较于两种传统方法而言, HMGCN相较于PALE, 在Macro Precision、Macro Recall和Macro F1分数上分别提高了14.46%、9.82%和9.47%; 对于GCN来说, 本文在Macro Precision、Macro Recall和Macro F1上分别提高了18.46%、13.32%和13.63%。由此看来, HMGCN在大规模社交网络上具备良好的性能。

在Weibo-Douban数据集(即小规模数据集)上, 本文提出的HMGCN相比于基准模型取得了更有竞争力的结果, 与MGCN相比, 在Macro Precision、Macro Recall和Macro F1分数上, 分别提高了18.49%、15.20%和17.92%, 与PALE、GCN以及CrossMNA相比在Macro Precision、Macro-Recall和Macro F1分数上, 提升也十分明显。对于MGCN来说, 虽然使用超图卷积来聚合多粒度信息, 但是在网络数据不足的情况下无法挖掘深层网络结构信息, 而HMGCN使用K近邻算法来聚合节点邻域特征, 对网络原始结构数据进行增强, 有效解决冷启动问题, GCN仅关注网络表层拓扑结构信息, 本文提出的HMGCN在结合网络表层特征的基础上, 通过网络同质性来进一步引导节点特征传播以及聚合过程, PALE则严重依赖于能够观察到的成对关系来进行节点表示学习, 不关注网络结构深层信息, CrossMNA使用网络间向量优化网络内表示学习效果, 但其不注重于网络内不同信息影响。这几种基线模型在小批量数据集上的效果都比较相近, 小批量数据集中可供模型提取的信息较少, 因此本文提出的HMGCN在社区化网络的基础上, 结合同质度矩阵深度挖掘社区内部信息, 并以此适应多尺度的社交网络, 实现高性能的跨网络用户对齐。

3.6 消融实验分析

HMGCN使用K近邻矩阵促进网络拓扑特征和节点特征融合, 以提升表示学习效果; 通过结合同质度矩阵对后续图卷积进行引导, 以达到挖掘社区内部深层结构信息的作用, 在本节中将分别对比模型在不同模块影响下图卷积网络的实验效果。

HMGCN_a为消除K近邻矩阵模块的模型, HMGCN_b为消除K近邻矩阵以及同质度矩阵引导模块的模型, 图6为其在ArnetMiner-LinkedIn数据

集上的实验结果,图7为其在 Weibo-Douban 数据集上的实验结果。

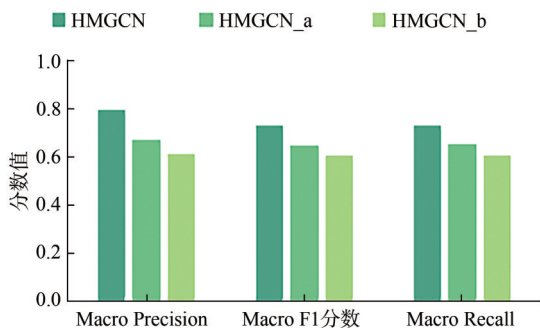


图6 ArnetMiner-LinkedIn数据集消融实验结果

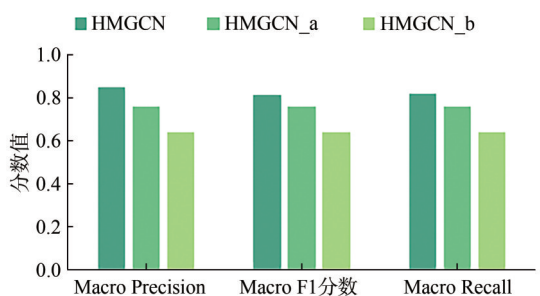


图7 Weibo-Douban数据集消融实验结果

由实验结果可得知, HMGCN相较于HMGCN_a与HMGCN_b在Macro Precision、Macro Recall和Macro F1分数上平均提升接近12%、7%和7%。相较于HMGCN_a而言, HMGCN以K近邻矩阵对网络原始数据进行增强,并在图卷积过程中结合节点特征实现优化图表示学习的效果,而与HMGCN_b相比, HMGCN中使用同质度矩阵引导图卷积过程,以实现网络节点邻域特征的传播与聚合,从而极大增强了网络对齐任务的效能。HMGCN不仅有效结合网络节点特征增强与网络同质性引导,并通过社区化结构处理社交媒体产生的多尺度社交网络,更能将其结果应用于不同的下游任务中。

3.7 模型鲁棒性分析

本节旨在通过控制锚链接数量,分析其对模型鲁棒性的影响,深入研究HMGCN在多尺度网络对齐中的性能表现。实验将基于20%、40%、60%、80%的锚链接比例进行,通过改变网络中数据组成来观察模型对监督信息的敏感性,以及在不同情况下的鲁棒性表现。图8为其在ArnetMiner-LinkedIn数据集上的鲁棒性实验结果,图9为其在Weibo-Douban数据集上的鲁棒性实验结果。

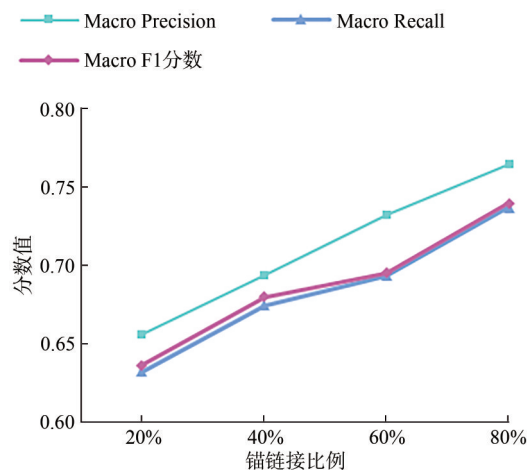


图8 ArnetMiner-LinkedIn数据集的鲁棒性实验结果

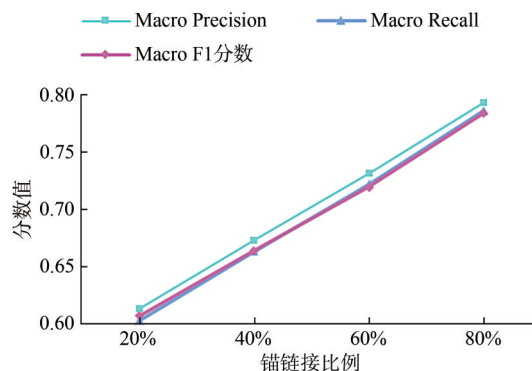


图9 Weibo-Douban数据集的鲁棒性实验结果

HMGCN在不同锚链接比例下都有着较为稳定的性能,即使是在可见锚链接比例处于20%的情况下,模型实现网络对齐任务的各项评价指标也与基线模型相差不大。除此之外,多尺度网络之间存在的差异性也会影响网络对齐模型的稳定性。ArnetMiner-LinkedIn数据集数据量大、网络结构较为复杂,锚链接数量与整体数据量相比差异较大。相较而言Weibo-Douban数据集数据量较小,网络结构较为清晰,锚链接数量占据整体数据量15%左右。因此,由图7和图8对比可知,ArnetMiner-LinkedIn数据集为模型提供了足量的网络结构信息,在不同锚链接比例下的各项评价指标下降较为平缓。而Weibo-Douban数据集中锚链接比例占比较大,即监督信息敏感性较高,所以在不同锚链接比例下的各项评价指标下降趋势明显,多尺度网络之间的差异性也是网络对齐任务中不可忽视的影响因素。

3.8 社区尺度敏感性分析

本节对HMGCN参数敏感性进行分析,对模型效果起主要影响的参数为社区大小,社区稀疏程度

会影响最终图卷积效果。本文控制学习率及嵌入维度不变,将社区范围大小作为变量来验证模型对不同尺度网络的卷积效果。对于 ArnetMiner-LinkedIn 数据集,本文设置第一组实验社区范围控制为 $N_{\min}=100$ 、 $N_{\max}=15\,000$;第二组实验社区范围控制为 $N_{\min}=1\,000$ 、 $N_{\max}=7\,500$;第三组实验社区范围控制为 $N_{\min}=1\,000$ 、 $N_{\max}=30\,000$ 。对于 Weibo-Douban 数据集,本文第一组实验社区范围控制为 $N_{\min}=10$ 、 $N_{\max}=1\,500$;第二组实验社区范围控制为 $N_{\min}=100$ 、 $N_{\max}=750$;第三组实验社区范围控制为 $N_{\min}=100$ 、 $N_{\max}=3\,000$ 。将两个数据集中的第一组实验中范围最小值缩小是为了考虑模型在网络结构信息不足情况下,能否提取足够网络信息,以实现网络对齐任务,其余两组实验则是为了验证模型在不同参数下的敏感度。在控制社区范围参数情况下,图10为 HMGCN 在 ArnetMiner-LinkedIn 数据集上的参数敏感性实验结果,图11为 HMGCN 在 Weibo-Douban 数据集上的参数敏感性实验结果。

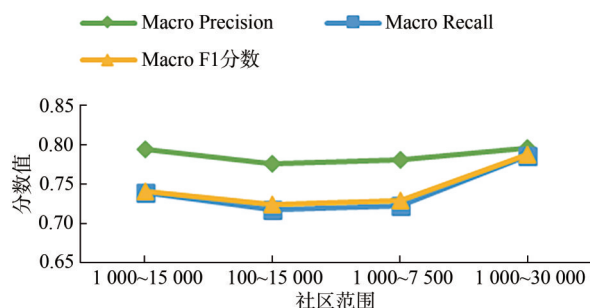


图10 ArnetMiner-LinkedIn数据集的参数敏感性实验结果

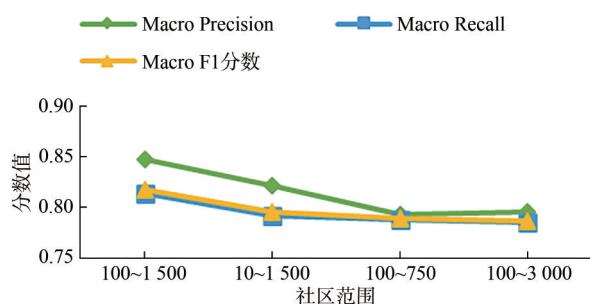


图11 Weibo-Douban数据集的参数敏感性实验结果

根据图9、图10实验结果分析可知,在网络结构信息不足的情况, HMGCN 能够提取足够网络信息以实现网络对齐任务;根据模型在不同参数下实验结果可知, HMGCN 可适应不同尺度网络下的网络对齐任务,以此解决网络对齐任务中网络规模的限制。

4 结束语

针对网络对齐领域存在的网络结构数据不足,以及处理不同规模社交网络的存储限制等问题,本文提出了一种结合同质度的多层图卷积网络和多尺度网络对齐框架。结合网络邻域特征对网络结构数据进行数据增强,分析网络内的同质性关系引导图卷积过程,并优化网络结构和节点特征的融合。并且本文通过社区发现算法和向量空间两阶段协调机制构造网络对齐框架,以适应多尺度网络。当前的人工智能系统在不同层次上都依赖大量的样本训练完成有监督的学习,真正的通用智能会在经验和知识积累的基础上灵活地进行无监督学习^[30],但本文提出的模型以及框架仍依赖于有效的监督信息,除此之外越来越多社交媒体平台提供诸多可见的用户,如何利用可见信息以及挖掘深层网络结构信息也是网络对齐领域研究者应该考虑的问题。因此未来的工作将探索结合用户属性以及网络结构的多尺度网络对齐问题,以及将本文的框架扩展到异质性社交网络对齐任务和动态网络对齐任务中。

参考文献:

- [1] CAO X Z, YU Y. BASS: A bootstrapping approach for aligning heterogeneous social networks[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2016: 459-475.
- [2] PENG C, XU K, WANG F, et al. Predicting information diffusion initiated from multiple sources in online social networks[C]//Proceedings of the 2013 Sixth International Symposium on Computational Intelligence and Design. Piscataway: IEEE Press, 2013: 96-99.
- [3] AHMAD M A, BORBORA Z, SRIVASTAVA J, et al. Link prediction across multiple social networks[C]//Proceedings of the 2010 IEEE International Conference on Data Mining Workshops. Piscataway: IEEE Press, 2010: 911-918.
- [4] MAN T, SHEN H W, HUANG J M, et al. Context-adaptive matrix factorization for multi-context recommendation[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 901-910.
- [5] TANG J L, GAO H J, LIU H, et al. eTrust: understanding trust evolution in an online world[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 253-261.
- [6] TAN S L, GUAN Z Y, CAI D, et al. Mapping users across networks by manifold alignment on hypergraph[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014, 28(1): 159-165.
- [7] IOFCIU T, FANKHAUSER P, ABEL F, et al. Identifying users across social tagging systems[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2021, 5(1): 522-525.
- [8] ZHAO Y, LIU Y, GUO X Y, et al. User naming conventions mapping

- learning for social network alignment[C]//Proceedings of 2021 13th International Conference on Computer and Automation Engineering (ICCAE). Piscataway: IEEE Press, 2021: 36-42.
- [9] MALHOTRA A, TOTTI L, MEIRA JR. W, et al. Studying user footprints in different online social networks[C]//Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE Press, 2012: 1065-1070.
- [10] RIEDERER C, KIM Y, CHAINTREAU A, et al. Linking users across domains with location data: theory and validation[C]//Proceedings of the 25th International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. New York: ACM Press, 2016: 707-719.
- [11] LIU S Y, WANG S H, ZHU F D, et al. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014: 51-62.
- [12] LIU J, ZHANG F, SONG X Y, et al. What's in a Name? an unsupervised approach to link users across communities[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2013: 495-504.
- [13] ZHANG H, KAN M Y, LIU Y, et al. Online social network profile linkage[C]//Proceedings of the 10th Asia Information Retrieval Societies Conference on Information Retrieval Technology. Cham: Springer International Publishing, 2014: 197-208.
- [14] ZHAO Y, LIU Y, GUO X Y, et al. User Naming conventions mapping learning for social network alignment[C]//Proceedings of the 2021 13th International Conference on Computer and Automation Engineering (ICCAE). Piscataway: IEEE Press, 2021: 36-42.
- [15] ZHAN Q Y, ZHANG J W, YU P S. Integrated anchor and social link predictions across multiple social networks[J]. Knowledge and Information Systems, 2019, 60(1): 303-326.
- [16] ZHANG J W, YU P S, ZHOU Z H. Meta-path based multi-network collective link prediction[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 1286-1295.
- [17] ZAFARANI R, LIU H. Connecting users across social media sites: a behavioral-modeling approach[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 41-49.
- [18] KONG X N, ZHANG J W, YU P S. Inferring anchor links across multiple heterogeneous social networks[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 179-188.
- [19] MAN T, SHEN H, LIU S, et al. Predict anchor links across social networks via an embedding approach[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2016: 1823-1829.
- [20] LI X X, CAO Y N, LI Q, et al. RLINK: Deep reinforcement learning for user identity linkage[J]. World Wide Web, 2021, 24(1): 85-103.
- [21] CHEN H X, YIN H Z, SUN X G, et al. Multi-level graph convolutional networks for cross-platform anchor link prediction[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2020: 1503-1511.
- [22] CHU X K, FAN X X, YAO D, et al. Cross-network embedding for multi-network alignment[C]//Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 273-284.
- [23] SAXENA S, CHANDRA J. SAlign: a graph neural attention framework for aligning structurally heterogeneous networks[J]. Journal of Artificial Intelligence Research, 2023, 77: 949-969.
- [24] LI Y, CUI H Y, LIU H L, et al. Triple-layer attention mechanism-based network embedding approach for anchor link identification across social networks[J]. Neural Computing and Applications, 2022, 34(4): 2811-2829.
- [25] LU M L, DAI Y L, ZHANG Z Q. Social network alignment: a bi-layer graph attention neural networks based method[J]. Applied Intelligence, 2022, 52(14): 16310-16333.
- [26] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): 10008.
- [27] ZHANG Y T, TANG J, YANG Z L, et al. COSNET: connecting heterogeneous social networks with local and global consistency[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015: 1485-1494.
- [28] CHEN B Y, CHEN X L. MAUIL: Multilevel attribute embedding for semi-supervised user identity linkage[J]. Information Sciences, 2022, 593: 527-545.
- [29] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 3844-3852.
- [30] 郑南宁. 人工智能新时代[J]. 智能科学与技术学报, 2019, 1(1): 1-3.
- ZHENG N N. The new era of artificial intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 1-3.

[作者简介]



崔家豪 (2000-), 男, 西北民族大学硕士生, 主要研究方向为社交网络对齐。



江涛 (1983-), 男, 博士, 西北民族大学教授, 主要研究方向为自然语言处理、社会计算和舆情分析。



徐梦瑶 (2000-), 女, 西北民族大学硕士生, 主要研究方向为谣言抑制。