DOI:10.11918/202501037

异质网络中融合多种语义关系的高效社区搜索方法

魏金阳[1],周丽华[1],王丽珍[1,2]

(1.云南大学 信息学院,昆明 650500;2.云南大学 滇池学院,昆明 650228)

摘 要:为解决异质信息网络中现有社区搜索方法存在的局限性,本文提出了一种融合多种语义关系的异质信息网络社区搜索方法,采用高效的"离线学习-在线搜索"策略,其核心在于:利用语义注意力机制自适应学习不同元路径对目标社区凝聚性的权重贡献,精准量化语义差异;再结合网络结构与节点属性特征度量节点相关性,定位社区成员。离线阶段预训练节点-社区关联模型,生成节点归属各类社区的概率分布向量;在线阶段基于预计算结果快速响应社区搜索。此策略既可保持学习模型的灵活性,有效捕捉异质网络语义与属性,又将主要计算负担置于离线阶段,显著提升查询效率,尤其适用于高频场景。在多个真实数据集上的验证实验表明,本方法在社区有效性(语义相关性、结构凝聚性、属性一致性)和查询效率上均显著优于现有主流方法。

关键词:异质信息网络;社区搜索;多种语义关系;离线学习;在线搜索

中图分类号:TP301 **文献标志码**:A **文章编号**:0367-6234(2026)01-0106-13

Efficient community search with multiple semantic relationships in heterogeneous information networks

WEI Jinyang[1], ZHOU Lihua[1], WANG Lizhen [1,2]

(1. School of Information Science & Engineering, Yunnan University, Kunming 650500, China;
2. School of Dianchi, Yunnan University, Kunming 650228, China)

Abstract: To address key limitations in existing community search methods for heterogeneous information networks (HINs), this paper proposes a community search method for HINs that integrates multiple semantic relationships, employing an efficient "offline learning, online search" strategy. Its core lies in: adaptively learning the weight contributions of different meta-paths to the target community cohesiveness using a semantic attention mechanism to precisely quantify semantic differences; and subsequently measuring node relevance by combining network structure and node attribute features to locate community members. In the offline phase, a node-community association model is pre-trained to generate probability distribution vectors indicating node affiliation across various communities. In the online phase, community search is rapidly responded to based on precomputed results. This strategy maintains the flexibility of the learning model to effectively capture heterogeneous network semantics and attributes, while shifting the main computational burden to the offline phase, significantly improving query efficiency, making it particularly suitable for high-query-frequency scenarios. Experiments on multiple real-world HIN datasets demonstrate that our method significantly outperforms existing mainstream methods in both community effectiveness (semantic relevance, structural cohesiveness, attribute consistency) and query efficiency.

Keywords: heterogeneous information networks; community search; multiple semantic relationships; offline learning; online query

真实世界的关系可以抽象为各种信息网络,如书目网络、社交网络、蛋白质网络等。社区搜索(community search,CS)[1]是从网络中找到包含查询节点的凝聚子图。由于其高度个性化和广泛的应用价值(例如事件组织、传染病监控、产品推广、舆情调控等),使其成为网络分析任务中的一个重要研究方向。

异质信息网络(heterogeneous information network,HIN)[2-3]是一种包含多种类型对象和连接关系的网络,其中不同类型的对象和连接关系从不同的维度刻画了网络的语义,可以更加完整自然地对现实世界的网络数据进行建模。因此,基于HIN的CS

收稿日期:2025-01-15;**录用日期**:2025-06-19;**网络首发日期**:2025-08-29
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20250829.0926.012
基金项目:国家自然科学基金(62562060,62062066,61762090,62276227);云南省基础研究计划重点项目(202201AS070015);云南省智能系统与计算重点实验室项目(202405AV340009)
作者简介:魏金阳(1998—),男,硕士研究生;周丽华(1968—),女,教授,博士生导师;王丽珍(1962—),女,教授,博士生导师
通信作者:周丽华,lhzhou@ynu.edu.cn

能够搜索到类型更为丰富的社区[4-5]，如(k,$\mathcal{P}$)-core[6]，(k, $\mathcal{P}$)-truss[7]，Butterfly-core[8]，Significance-(α,β)-community[9]。这些社区语义丰富，易于解释。然而，HIN 中丰富的语义信息及节点和连边的多样性也给 CS 带来了挑战。首先，不同类型对象之间的连接存在不同的语义，可能具有不同的形成机制，同时也会相互关联和相互影响，导致各自在 CS 中的作用也不同。因此，区分不同的语义关系，即区分各种对象和连接及其在 CS 中的作用是必要的。其次，不同类型的对象可以通过不同类型的关系连接在一起，同种类型的对象也可以通过不同的方式组织在一起，因此，在异质信息网络中 CS 的搜索空间更大，计算更复杂，如何设计高效的搜索算法以满足实时、可扩展的要求也是 CS 的一个关键问题。

现有的 CS 方法可以分为基于规则的方法和基于学习的方法。基于规则的方法可以在给定的预定义结构约束(如(k,$\mathcal{P}$)-core[6]、$k\mathcal{P}$-core[10]等)上发现满足特定条件的社区，这种方法找到的社区结构凝聚性较高，但是参数 k 的指定较为困难，较高的 k 可能找不到期望的社区，而较低的 k 可能导致返回的社区规模较大，使得社区查找不够灵活；基于学习的方法是通过模型训练来定位社区，可灵活地捕获到社区的结构和属性相似性(如 ICS-GNN[11]等)，但是，针对每个查询都需要重新训练模型会影响搜索的效率，且由于放宽了社区的结构限制(仅需要保持社区连通)，导致结果社区的凝聚性较低。另外，大部分基于学习的方法都是面向节点和边类型单一的同质网络设计的，难以应用于包含丰富语义的 HIN 中查询社区。

鉴于使用规则和学习模型的 CS 方法存在的问题及不同语义关系对目标社区重要性捕捉的必要性，本研究提出了异质信息网络中融合多种语义关系的社区搜索方法(heterogeneous network community search with multiple semantic relationships，HCSMS)及采用离线学习和在线搜索的高效搜索策略。HCSMS 是一种基于图神经网络(graph neural network，GNN)的 CS 方法，通过多条元路径来捕获多种语义关系，其中，不同语义关系对目标社区的重要性通过语义注意力模型[12]在不断的训练和迭代中得到学习，并在结合网络结构特征和节点属性特征的基础上，指导节点与查询节点相关性的衡量，以产生节点位于不同类别社区的概率。本研究将目标类型节点与查询节点属于同一类别社区的概率定义为节点评分，并通过一种贪心策略来寻找一定规模(社区内的节点数目)内具有最大评分总和的稠密连通子图作为目标社区，社区的评分总和最大，意味着社区成员与查询节点最相关。为了加强目标社区的结构凝聚性，本文进一步引入 k-core 模型来提高社区的凝聚性。

由于模型学习到目标类型节点属于不同类别社区的概率与查询节点无关，因此，离线学习策略可以预先学习到目标类型节点属于不同类别社区的概率。在线搜索阶段输入查询节点，并根据查询节点的类别从离线阶段训练好的模型上获得节点评分，然后综合结构凝聚性和社区大小约束定位社区。由于不需要针对每个查询节点重新训练模型来获得节点评分，因此社区搜索速度得到提升，且查询频率越高，查询效率提升越明显。另外，对于评分相差较大的两个相邻节点，他们属于相同社区的可能性相较于评分相近的节点会更小。基于这个观察，本文提出一种网络优化策略，将概率相差较大节点间的连边删除，从而减少在线搜索时遍历的时间，进一步提高社区定位的效率。

1　研究概况

当前社区搜索研究主要可分为基于规则和基于学习两类方法。

基于规则的社区搜索依赖预定义规则定位社区。这类方法通常使用预定义规则来查找社区，如(k,$\mathcal{P}$)-core[6]求指定元路径下社区内节点度大于 k；(k,$\mathcal{P}$)-truss[7]规定社区中每条边至少存在于 $k-2$ 个三角形中的最大子图[13]；$k\mathcal{P}$-core[10]在(k,$\mathcal{P}$)-core 基础上结合节点属性关键字，增强属性相似性捕获；Butterfly-Core [8]以两个查询节点为中心，基于 k-core[14]搜索构建蝴蝶状密集社区；Significance-(α,β)-community[9]基于二部图，通过节点参与度与边权重变化挖掘社区；基于主题的交互图和(k,l,η)-有影响力社区[15]，则致力于解决主题感知下的社区搜索问题。

基于学习的社区搜索通过模型训练度量节点相关性，突破规则限制。早期方法多基于随机游走，如 Guo 等[16]引入了社区的结构规则，提出属性核等方法；Zhao 等[17]采用不立即回访模型实现社区向量检索；Liu 等[18]以亲密度和孤立度综合评估节点关系。基于图神经网络的研究，如 Gao 和 Chen 等[11,19]提出的 ICS-GNN，将 CS 转化为二分类问题，通过候选子图训练提升效率。Jiang 等[20]提出的 QD-GNN，支持属性查询，优化检索条件。但上述方法多针对同质网络，难以适配异质信息网络(HIN)。近年面向 HIN 的研究中，ICSMIM[21]通过挖掘互信息最大节点集构建社区；VMCS-DGNN[22]在指定元路径训练模型，并利用 0/1 背包问题平衡结构与属性指标；

SNCS[23]融合拓扑和潜在特征,实现语义约束下的社区搜索,CS-DAHIN[24]研究动态异质网络的社区搜索。

本文所提 HCSMS 方法是一种基于学习的社区搜索方法,与现有研究的区别在于,HCSMS 以进一步捕获 HIN 中不同语义关系对目标社区的影响来发现社区,且 HCSMS 使用的查询框架可以不需要针对每轮查询重新训练模型,提高了搜索的效率。

2 基本概念及问题描述

2.1 基本概念

定义 1 异质信息网络(HIN)[2]。HIN 定义为一个具有对象类型的映射函数 $\varphi: V \rightarrow T$(节点→节点类型)和关系映射函数 $\phi: E \rightarrow R$(边→边类型)的图 $G=(V,E)$,节点和边均属于特定类型,如图 1(a)所示。

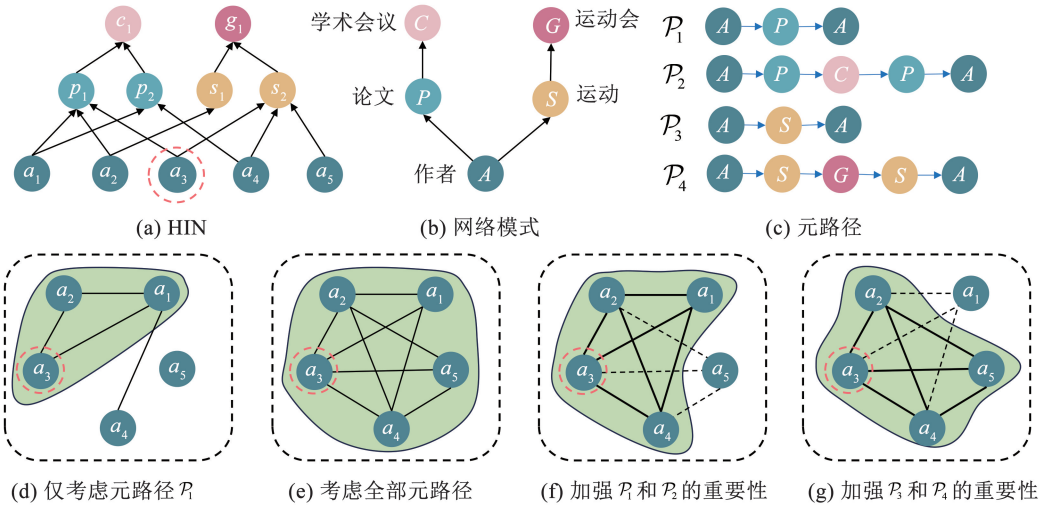


图 1 语义关系分析示意图

Fig. 1 Semantic relationship analysis diagram

定义 2 网络模式[6]。是定义在节点类型集合 T 和边类型集合 R 上的一个有向图 $S_G=(T,R)$,其是 HIN 的元描述,引导网络语义的探究,如图 1(b)所示。网络中的节点是网络模式的实例化,例如,a_1 是作者类型 A 的一个实例。

定义 3 元路径[2]。在网络模式中由边序列连接的节点类型序列$\mathcal{P}$,可表示为特定形式,其长度有相应定义,元路径能够描述节点间的语义关系,如 *APA* 描述论文合作关系,图 1(a)中,$a_1p_1a_3$ 为元路径 *APA* 的一个实例。

定义 4 $\mathcal{P}$-邻居。通过元路径$\mathcal{P}$与节点 i 连接的所有节点集合 $N_i^{\mathcal{P}}=\{v_1,v_2,\cdots\}$。其中,$i$ 的$\mathcal{P}$-邻居包含节点本身。

定义 5 $\mathcal{P}$-连通图。HIN 通过元路径$\mathcal{P}$诱导的一个同质图,该图中任意一个节点 i 的所有直接邻居均为 i 的$\mathcal{P}$-邻居。如图 1(d)描述的是图 1(a)中 HIN 的$(\mathcal{P}_1)$-连通图,图 1(e)描述的是图 1(a)中 HIN 的$(\mathcal{P}_1,\mathcal{P}_2,\mathcal{P}_3,\mathcal{P}_4)$-连通图。

2.2 问题描述

语义关系捕获不充分。部分方法仅依赖单条元路径刻画单一语义关系,难以全面捕捉网络复杂性;其他一些方法虽引入多条元路径描述多元语义,但未对其重要性进行差异化考量,导致社区质量受限。图 1(a)是一个包含作者学术和体育关系的 HIN,同个查询 a_3 基于不同元路径的处理方式将找到不同的社区,即图 1(d)、(e)、(f)和(g)。其中,图 1(d)仅基于单一语义导致社区较为稀疏,图 1(e)未差异化考量元路径导致社区联系不够紧密,图 1(f)和图 1(g)对不同的元路径进行区分融合而找到了不同的社区(学术相关和运动相关),易于解释且更符合实际情况。因此,如何量化不同语义关系的重要性并实现有效融合以提升社区搜索效果,仍是亟待探索的关键问题。

语义关系学习成本。现实世界中的 HIN 节点和边类型繁杂,需要大量的元路径描述,人为地指定元路径关系来查找社区成本极高。因此,如何从多元语义关系中自动学习并识别对目标社区最为关键的语义关系,成为 HIN 社区搜索领域亟待突破的核心问题。

搜索效率待提高。基于学习模型的 ICS-GNN,虽通过构建候选子图查询在一定程度上提升了效率,但每次查询均需建立候选子图训练模型,训练耗时久,在查询频率较高时,难以满足及时搜索的需求。

2.3　问题定义

给定一个 HIN $G=(V,E)$，一个查询节点 q，社区规模 s，G 中的元路径集合为 $M=\{\mathcal{P}_1,\cdots,\mathcal{P}_m\}$。本文的目标是为 q 找到一个社区 $C=(V_c,E_c)$，$q\in V_c$，对于任意一个节点 $v\in V_c$，有 $\varphi(v)=\varphi(q)$，且 V_c 中的节点之间要具有属性相似性。E_c 中的边是紧密的保证社区内聚结构，E_c 至少包含两种以上元路径连通图关系，且包含的这些元路径连通图关系具有相关性，保证社区是语义相关的。

3　社区搜索框架设计

为了找到属性相似、结构内聚和语义相关的社区，本文设计了社区搜索方法 HCSMS，其框架包含离线学习和在线搜索两个阶段，如图 2 所示。离线学习针对多种语义信息融合网络中的结构及节点属性，以计算各个节点属于不同类别社区的概率；在线查询时，输入查询节点并依据其标签类别，匹配离线阶段预计算的节点–社区归属概率（即节点评分），实现社区精准定位。节点评分反映了节点与查询节点的相关性，节点评分高，意味着节点与查询节点相关性高。

图 2 中输入的 HIN 包含 4 种节点类型（4 种形状）及 6 种连边类型，形状相同但颜色不同的节点表示其属于相同的节点类型，但标签类别不同，如节点 a 和 b 均为作者，但作者 a 的标签可能是人工智能，作者 b 的标签可能是数据库。节点的属性描述了节点的特征，每种类型节点的属性用属性特征矩阵 $\boldsymbol{H}$ 描述。

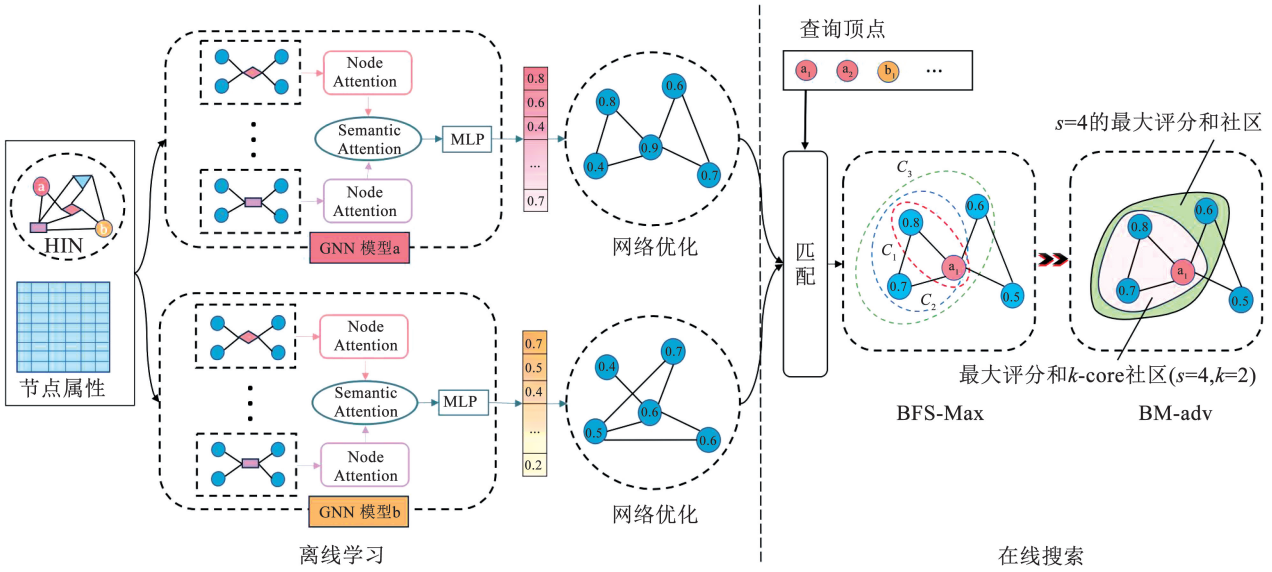


图 2　HCSMS 框架

Fig. 2　Framework of HCSMS

3.1　离线学习

离线学习阶段包含节点评分计算和网络优化两个模块。

3.1.1　节点评分计算

在图 2 中，圆形节点 a 和 b 的标签属于两种类别（两种颜色），需要分别训练两个 GNN 模型。两个 GNN 模型分别以节点 a 和节点 b 的标签为正例进行训练，训练结束后获得所有圆形节点属于两种社区类别的概率（评分）。

算法 1 概述了离线阶段的核心流程：基于节点标签类别训练 GNN 模型，计算节点对不同社区类别的归属评分。第 1 行的循环根据每种标签类别训练一个 GNN 模型，第 2 行初始化节点评分向量 $\boldsymbol{\delta}_i$ 为一个一维零向量，第 3 行循环使 GNN 模型收敛，第 4 行中的 $v.lable$ 表示节点 v 的标签，若 $v.lable$ 属于

算法 1　节点评分计算

输入：异质信息网络 $G=(V,E)$，节点属性特征矩阵 $\boldsymbol{H}$，
元路径集合 $M=\{\mathcal{P}_1,\cdots,\mathcal{P}_m\}$，标签类别 $L=\{\ell_1,\cdots,\ell_n\}$；

输出：节点评分 $\boldsymbol{\delta}=\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_n\}$

1　**for** $\ell_i\in L$ **do**
2　　初始化节点评分 $\boldsymbol{\delta}_i$
3　　**while** $loss(\ell_i,\boldsymbol{\delta}_i)\geqslant\gamma$ **do**
4　　**for** $v\in V$ and $v.lable$ is ℓ_i **do**
5　　**for** $\mathcal{P}_x\in M$ and $\mathcal{P}_x$.head is $\varphi(v)$ **do**
6　　通过节点级注意力学习嵌入 $\boldsymbol{z}_v^{\mathcal{P}_x}$//等式(1.1)
7　　通过语义级注意力学习嵌入 $\boldsymbol{z}_v$//等式(1.2)
8　　计算节点评分 $\boldsymbol{\delta}_i[v]=MLP(\boldsymbol{z}_v)$
9　**return** $\boldsymbol{\delta}=\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_n\}$

ℓ_i 类别，则由行 5 选择以 v 的类型为起始节点的元路径集合，第 5 行中的$\mathcal{P}_x.head$ 表示元路径$\mathcal{P}_x$ 的起始节点，第 6 行基于元路径$\mathcal{P}_x$ 以及节点级注意力机制计算节点 v 的节点级嵌入 $\boldsymbol{z}_v^{\mathcal{P}_x}$，聚合公式见式(1)。

$$\begin{cases}\boldsymbol{z}_v^{\mathcal{P}_x}=\sigma\left(\sum\limits_{j\in N_v^{\mathcal{P}_x}}\alpha_{vu}^{\mathcal{P}_x}\cdot\boldsymbol{h}_u\right)\\ \alpha_{vu}^{\mathcal{P}_x}=\dfrac{\exp\left(\sigma\left(\boldsymbol{a}_{\mathcal{P}_x}^{\mathrm{T}}\cdot[\boldsymbol{h}_v\,\|\,\boldsymbol{h}_u]\right)\right)}{\sum_{k\in N_v^{\mathcal{P}_x}}\exp\left(\sigma\left(\boldsymbol{a}_{\mathcal{P}_x}^{\mathrm{T}}\cdot[\boldsymbol{h}_v\,\|\,\boldsymbol{h}_k]\right)\right)}\end{cases}\tag{1}$$

式中：$\sigma(\cdot)$为非线性激活函数，$N_v^{\mathcal{P}_x}$ 为节点 v 在元路径$\mathcal{P}_x$ 下的元路径邻居集合，$\boldsymbol{h}_v$ 为节点 v 的属性特征矩阵，$\alpha_{vu}^{\mathcal{P}_x}$为在元路径$\mathcal{P}$下节点 v 对节点 u 的注意力系数，$\boldsymbol{a}_{\mathcal{P}_x}^{\mathrm{T}}$为在元路径$\mathcal{P}_x$ 下的注意力向量。

第 7 行使用语义级注意力聚合来自不同元路径的 $\boldsymbol{z}_v^{\mathcal{P}}$ 得到节点 v 的最终嵌入 $\boldsymbol{z}_v$，计算公式如下：

$$\begin{cases}\boldsymbol{z}_v=\sum\limits_{x=1}^{m}\boldsymbol{z}_v^{\mathcal{P}}\cdot\dfrac{\exp(w^{\mathcal{P}_x})}{\sum\limits_{y=1}^{m}\exp(w^{\mathcal{P}_y})}\\ w^{\mathcal{P}_x}=\dfrac{1}{|V|}\sum\limits_{v\in V}\boldsymbol{q}^{\mathrm{T}}\cdot\tanh(\boldsymbol{W}\cdot\boldsymbol{z}_v^{\mathcal{P}}+\boldsymbol{b})\end{cases}\tag{2}$$

式中：$\tanh(\cdot)$为激活函数，V 为顶点集合，$\boldsymbol{q}^{\mathrm{T}}$ 为语义注意力向量，m 表示元路径数量，$\boldsymbol{W}$ 为权重矩阵可学习参数，$\boldsymbol{b}$ 为偏置项，$w^{\mathcal{P}_x}$为元路径$\mathcal{P}_x$ 的注意力系数，用于衡量$\mathcal{P}_x$ 的重要性。

第 8 行将 $\boldsymbol{z}_v$ 通过多层感知机(MLP)计算节点 v 的评分 $\boldsymbol{\delta}_i[v]$。最后，计算完所有标签类别下的所有节点的评分，获得 $\boldsymbol{\delta}=\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_n\}$，完成算法。

模型训练的损失计算如等式为

$$loss(\ell_i,\boldsymbol{\delta}_i)=\sum_{v.\,lable=\ell_i}-y_v\cdot\ln(\boldsymbol{\delta}_i[v])-(1-y_v)\cdot\ln(1-\boldsymbol{\delta}_i[v])\tag{3}$$

式中：$\boldsymbol{\delta}_i[v]$表示节点 v 属于第 i 类标签社区的预测概率，y_v 表示节点 v 属于第 i 类标签社区的真实概率，若 $v.lable=\ell_i$，则 $y_v=1$，否则 $y_v=0$。

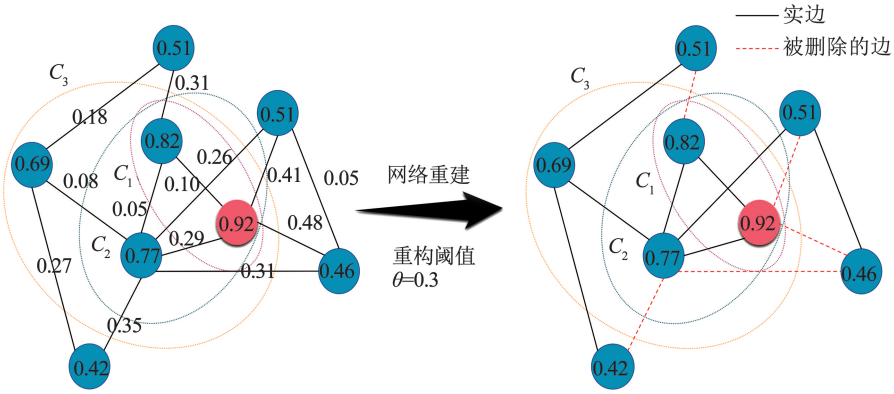
通过对核心步骤的分析，算法的时间复杂度可以表示为 $O(n_L\cdot n_\gamma\cdot n_M\cdot n\cdot d)$，由于标签数量 n_L，元路径数 n_M，嵌入维度 d 均为常数级，因此，时间复杂度可简化为 $O(n\cdot n_\gamma)$，其中，n_γ 为最大迭代次数，n 为节点数。空间复杂度主要由节点特征和嵌入维度决定，且语义级融合与元路径数量有关，因此算法 1 的空间复杂度可以表示为 $O(n_M\cdot n\cdot(D+d))$，由于节点特征维度 D 要远大于嵌入维度 d 以及元路径数量 n_M，因此可以简化为 $O(n\cdot D)$。

3.1.2 网络优化

评分相差较大的两个邻居节点属于相同社区的可能性比评分相近的邻居节点小。基于这个观察，本文提出一种网络优化策略，将网络中评分相差较大的节点间的连边删除，从而减少在线搜索时节点遍历的时间，进而提高社区定位的效率。

图 3(a)是圆形类型节点与其元路径邻居的连接关系图，圆形节点内的数值是算法 1 计算的节点属于蓝色标签类别的评分，连边上的数值表示边的两个端点的评分差的绝对值，设置评分差阈值 $\theta=3$，则图 3(a)中有 4 条边被删除，优化后的网络如图 3(b)所示。以定位社区 C_3 为例，不经过网络优化前，需要遍历 23 次，优化后，只需要遍历 14 次，有效提高了社区定位的效率。

网络优化过程如算法 2 所示，其中第 1 行循环表明每种标签类别的网络都要优化，第 2 行初始化优化图的节点集和边集，其中 $N_u^{u.\mathcal{P}}$ 表示 u 在所有元路径下元路径邻居的集合，第 3 行和第 4 行遍历优化图节点集中的所有节点及其邻居，第 5 行删除评分差小于阈值 θ 的边。该算法的时间和空间复杂度主要取决于节点的数量 n，其时间复杂度可以表示为 $O(n^2)$，空间复杂度为 $O(n)$。



(a) 圆形类型节点与其元路径邻居连线关系图　　(b) 优化后网络

图 3　网络优化

Fig. 3　Network optimization

算法 2　网络优化

输入：异质信息网络 $G=(V,E)$，节点评分 $\boldsymbol{\delta}=\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_n\}$，评分差阈值 θ；

输出：优化图集合 $\{G'_1,G'_2,\cdots,G'_n\}$

1　**for** $\boldsymbol{\delta}_i\in\{\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_n\}$ do

2　　初始化：$V'_i=\{v\mid\varphi(v)=\varphi(u),\ u.lable=l_i\}$

$E'_i=\{(u,v)\mid u,v\in V'_i,v\in N_u^{u.\mathcal{P}}\}$

3　　**for** $u\in V'_i$ **do**

4　　**for** $(u,v)\in E'_i$ **do**

5　　**if** $|\boldsymbol{\delta}_i[u]-\boldsymbol{\delta}_i[v]|>\theta$ **then**

6　　$E'_i=E'_i\setminus\{(u,v)\}$

7　　优化完成 G'_i

8　**return** $\{G'_1,G'_2,\cdots,G'_n\}$

3.2　在线搜索

在线搜索输入查询节点，根据查询节点的标签类别匹配离线阶段学习好的与查询节点属于同类别社区的概率，然后定位目标社区。目标社区定义为具有最大综合评分的稠密连通子图。

定义 6　最大评分和社区。给定一个 HIN $G=(V,E)$，查询节点 q，社区规模 s 以及节点评分 $\boldsymbol{\delta}$，最大评分和社区是 G 中满足如下条件的一个连通子图 $C=(V_c,E_c,\boldsymbol{\delta}_c)$：

(1) 查询节点 $q\in V_c$，且 V_c 中的任意节点的类型与 q 的类型相同，即 $\varphi(v_i)=\varphi(v_q)$；

(2) $|V_c|\leqslant\varepsilon$，且 C 保持连通；

(3) C 中节点评分总和 $\sum_{v\in V_c}\boldsymbol{\delta}_c[v]$ 是最大的。

条件(1)要求查询节点包含在目标社区中，社区中的节点类型与查询节点保持一致；条件(2)对社区的规模进行了限制，且社区必须是连通的。条件(3)要求社区内节点评分总和最大，即保证社区内所有节点的标签类别与查询节点的标签类别是最相关的，意味着属性相似性最高。

为了搜索最大评分和社区，提出了 BFS-Max 算法。BFS-Max 的主要思想是从当前社区出发，每次仅遍历当前社区内节点的所有邻居，将社区外评分最大的邻居节点加入社区。当前社区初始化为查询节点。图 2 中在线搜索对 BFS-Max 算法实现过程进行了演示，第 1 轮迭代后，社区更新为 C_1，然后遍历 C_1 的所有邻居，搜索评分最大的节点加入社区，第 2 轮迭代后，社区更新为 C_2。重复这个过程，直至社区规模满足要求。

算法 3 的伪代码描述了 BFS-Max 的实现过程。第 1 行根据查询节点 q 的标签的类别 l_i 选择优化图 $G'_i=(V'_i,E'_i,\boldsymbol{\delta}_i)$，第 2 行初始化目标社区中的节点和边，第 3 行循环找到满足社区规模的社区，第 4 行选择当前社区内所有节点的所有不在社区内的具有最大评分的元路径邻居，第 5 行将选择的邻居及相应的边加入社区，完成搜索。

算法 3　BFS-Max

输入：优化图集合 $\{G'_1,G'_2,\cdots,G'_n\}$，查询节点 q，社区规模 s；

输出：最大评分和社区 $C=(V_c,E_c,\boldsymbol{\delta}_c)$

1　根据 q 标签类别选择优化图 $G'_i=(V'_i,E'_i,\boldsymbol{\delta}_i)$

2　初始化：$V_c=\{q\}$；$E_c=null$

3　　**while** $|V_c|<s$ **do**

4　　$v=\arg_{u,x}\max\boldsymbol{\delta}_i[u]$，$u\notin V_c$，$u\in N_x^{x.\mathcal{P}}$，$x\in V_c$

5　　$V_c=V_c\cup\{v\}$，$E_c=E_c\cup\{(x,u)\}$

6　**return** $C=(V_c,E_c,\boldsymbol{\delta}_c)$

该算法的时间复杂度可以表示为 $O(n_L\cdot s\cdot n'\cdot e)$，其中标签数量 n_L 和社区规模为常数级 s，图优化后节点数量 n' 远大于节点的边数 e，因此，可以简化为 $O(n')$；空间复杂度表示为 $O(s+n'+e)$，同理可简化为 $O(n')$。

由于一定规模内具有最大评分和的连通子图的结构凝聚性可能不高，因此本文提出 BFS-Max 的增强算法 BM-adv，在挖掘到最大概率和的连通子图后，进一步通过 k-core 模型来挖掘凝聚性强的社区。如图 2 在线搜索中，展示了 $s=4,k=2$ 找到的最大评分和社区以及最大评分和 k-core 社区。BM-adv 的伪代码如算法 4 所示。

第 1 行通过 BFS-Max 找到社区，第 2 行的循环确保社区内所有节点的度大于 k，第 3,4 行搜索不满足 k-core，第 5,6 行删除不满足 k-core 的节点及在社区内的连边。

该算法的时间和空间复杂度取决于优化后的图的节点数量 n'，其复杂性与算法 3 类似，因此，其时间和空间复杂度均为 $O(n')$。

算法 4　BM-adv

输入：优化图集合 $\{G'_1,G'_2,\cdots,G'_n\}$，查询节点 q，社区规模 s，k-*core* 的 k；

输出：最终社区 $C'=(V'_c,E'_c,\boldsymbol{\delta}'_c)$

1　$C'=BFS\text{-}Max(G',q,s)$

2　**while** $\exists v\in V'_c,|N_v^{v.\mathcal{P}}|<k$ **do**

3　　**for** $v\in V'_c$ **do**

4　　**if** $|N_v^{v.\mathcal{P}}|<k$ **then**

5　　$V'_c=V'_c\setminus\{v\}$

6　　$E'_c=E'_c\setminus\{(v,u)\mid u\in V'_c,u=N_v^{v.\mathcal{P}}\}$

7　**return** $C'=(V'_c,E'_c,\boldsymbol{\delta}'_c)$

4　实验及分析

本文在 3 个异质信息网络上进行了大量实验，以进行如下验证。

W1：区分不同语义关系的重要性是否能够找到

更高质量的社区。

W2:使用 GNN 模型学习网络结构、节点属性以及语义信息对节点相关性进行综合衡量,验证是否能够提高社区的质量。

W3:离线学习和在线查询的框架以及网络优化策略是否能够提高社区搜索的效率。

4.1 实验设置

4.1.1 数据集

本文使用 3 个 HIN 数据集进行实验,数据集的相关信息如表 1 所示。

表 1 数据集相关信息

Tab. 1 Dataset related information

数据集	节点类型	节点数	边类型	边数	元路径
ACM	* Paper(P)	4 025	P-P	9 744	*PSP*
	Author(A)	7 167	P-A	13 407	*PAP*
	Subject(S)	60	P-S	4 019	
DBLP	* Author (A)	4 057	A-P	19 645	*APA*
	Paper(P)	14 328	P-C	14 328	*APCPA*
	Conf(C)	20	P-T	85 810	*APTPA*
	Term(T)	7 723			
IMDB	* Movie(M)	3 550	M-A	12 831	*MAM*
	Actor(A)	5 432	M-D	4 181	*MDM*
	Director(D)	2 083			

ACM 数据集[12]的论文、作者、主题 3 种类型节点来自数据库、无线通信和数据挖掘领域的论文发表记录。实验过程中查询节点从论文中选择,论文属性针对 1 803 个关键词进行了 one-hot 编码,维度为 1 803,论文的标签类别有 3 种。

DBLP 数据集[25]包含数据挖掘、人工智能、计算机视觉和自然语言处理领域相关的作者、论文、会议、术语 4 种类型节点,实验过程中查询节点从作者中选择,作者属性是 one-hot 编码,维度为 2 000,作者的标签类别有 4 种。

IMDB 数据集[12]包含电影、演员、导演 3 种类型的节点,实验过程中查询节点从电影中选择,其属性是电影情节的 one-hot 编码,维度为 1 007,电影的标签类别有 3 种。

4.1.2 对比算法

本文选择了几种基线算法作为对比算法,对比算法的描述如下。

4.1.2.1 ICS-GNN[11]

ICS-GNN 是基于 GNN 的同质网络社区搜索方法,其主要思想是在大图上建立社区候选子图训练 GNN 模型,并通过用户指导的交互方式在子图上定位社区。ICS-GNN 设计了 BFS-swap、Greedy-G 和 Greedy-T 三种社区搜索算法。与 ICS-GNN 对比分析的目的,是验证 HCSMS 在异质网络中挖掘语义信息进行社区搜索的有效性。

1) BFS-swap. 该算法首先在候选子图上通过 BFS(广度优先遍历)方式找到一个规模为 s 的社区,然后通过节点交换的方式,将社区外评分大的节点与社区内评分低的节点进行交换,从而保证社区评分最大。

2) Greedy-G. 该算法是为应对当查询节点位于社区边缘时定位最大评分社区的困难,通过节点与社区的最短距离,结合节点评分综合计算收益以挖掘社区。

3) Greedy-T. 该算法是 Greedy-G 的简化版,使用节点与查询节点的距离来代替最短路径的计算。

4.1.2.2 $(k,\mathcal{P})$-core[6]

$(k,\mathcal{P})$-core 是一种仅考虑结构内聚性挖掘社区的模型,该模型的搜索算法 FastBCore 是在指定元路径下寻找满足节点度大于 k,且结构连接紧密的凝聚子图。与 FastBCore 的比较是为验证所提社区搜索方法协同网络拓扑结构和节点属性,在社区定位上的有效性。

4.1.2.3 ICSMIM[21]

ICSMIM 是基于 GNN 的异质网络的社区搜索方法。该方法使用互信息来描述节点之间的相关性,通过社区搜索算法 MI-Max 挖掘异质网络中互信息最大的节点以构建目标社区。与 ICSMIM 对比的目的,是验证本文所提的社区度量指标相较于互信息的有效性。

4.1.2.4 VMCS-DGNN[22]

该方法是基于解耦图神经网络的异质网络社区搜索方法,其是在大图上建立候选子图来训练 GNN 模型,然后基于 0/1 背包问题对社区属性和结构凝聚性指标进行平衡来寻找社区。VMCS-DGNN 需要指定元路径来搜索社区,与 VMCS-DGNN 对比的目的,是验证多种语义关系融合寻找社区及离线学习和在线搜索框架的有效性。

4.1.3 评价指标

本文选取社区精确度(Precision)[11]、社区 F1-score[26]和社区密度[27]衡量社区质量,使用查询时间(Time)衡量社区查询速率。

1)社区精确度

社区精确度定义为结果社区中,标签与查询节点一致的节点占所有节点的比例,精确率越高说明模型性能越好。

2)社区 F1-score

F1-score 是准确率和召回率的调和平均值,用于衡量搜索社区与真实社区的接近程度,F1-score

越高说明结果越精确。

3)社区密度

社区密度计算公式为 $m/[n\times(n-1)/2]$，其中 m 为社区内部的总边数，n 为社区内节点的个数。社区密度用于衡量社区凝聚程度，社区密度越高，说明社区结构越紧密。

4)社区搜索效率

社区搜索效率是指完成社区搜索过程需要的时间。查询需要的时间越少，查询效率越高。

4.1.4　实现细节

本文随机初始化可学习参数，并通过 Adam 对模型进行优化。在训练过程中，设置学习率为 0.001，正则化参数为 0.001，衰退率为 0.6，隐藏层数为 2，注意力头数为 4，语义级注意力可学习参数 $\boldsymbol{q}$ 的维度为 128，最终嵌入维度为 2，迭代次数为 200。在所有对比算法中，为了保证公平性，在无特别情况说明时，默认社区规模为 30，并按照同一组查询节点进行社区搜索。对于算法 ICS-GNN，本文在目标类型节点(表 1 中带 * 的类型)的元路径诱导连通图中进行搜索，模型训练参数设置与原文一致。对于算法 FastBcore，社区规模会受到社区结构 $(k,\mathcal{P})$-core 中 k 值的影响，本文选取了社区规模可以接近 30 的 k 值作为 $(k,\mathcal{P})$-core 的 k。本文算法基于 Python3.8 实现，运行环境为 AMD Ryzen 75800H CPU(3.20 GHz)、16 GB 内存及 GeForce RTX 3080 GPU，模型训练通过 PyTorch 框架完成。

4.2　模型有效性分析

4.2.1　社区质量对比分析

表 2 给出了不同社区搜索算法在不同数据集上的精确度和 F1-score 以及社区密度值，可以看到，HCSMS 通过 BFS-Max 算法找到的社区在 3 个数据集上均取得了最高的精确度和 F1-score，BM-adv 算法在 3 个数据集上取得了最高的社区密度值，说明 HCSMS 挖掘到的社区的质量优于基线方法。

表 2　不同社区搜索算法的社区质量对比

Tab. 2　Comparison of community quality among different community search algorithms

数据集	指标	FastBCore	VMCS-DGNN	ICS-GNN			ICSMIM	HCSMS	
				BFS-swap	Greedy-T	Greedy-G	MI-Max	BFS-Max	BM-adv
ACM	Precision	0.430	0.954	0.885	0.890	0.905	0.952	**0.968**	0.945
	F1-score	0.420	0.920	0.855	0.850	0.885	0.920	**0.932**	0.920
	Density	0.747	0.672	0.664	0.652	0.662	0.668	0.672	**0.754**
DBLP	Precision	0.310	0.850	0.825	0.828	0.848	0.852	**0.956**	0.952
	F1-score	0.300	0.820	0.800	0.805	0.815	0.820	**0.920**	0.915
	Density	0.960	0.936	0.912	0.925	0.930	0.930	0.943	**0.963**
IMDB	Precision	0.425	0.592	0.515	0.535	0.564	0.570	**0.728**	0.720
	F1-score	0.400	0.582	0.506	0.535	0.550	0.562	**0.676**	0.664
	Density	0.670	0.620	0.640	0.590	0.612	0.633	0.642	**0.673**

FastBcore 找到的社区有较高的密度，但精确度、F1-score 低于其他几种社区搜索算法，这是因为 FastBcore 仅聚焦于结构凝聚性，既未融合节点属性信息，也没有捕获 HIN 中的多元语义关系。ICS-GNN 的 3 个社区定位算法中，Greedy-G 算法的精确度、F1-score 和社区密度值最高，但 Greedy-G 在 3 个数据集上的精确度、F1-score 和社区密度值均低于 MI-Max、BFS-Max、BM-adv 的值，而 MI-Max 的值比 BFS-Max、BM-adv 的低，说明基于本文所提方法的社区度量方式优于使用互信息度量，VMCS-DGNN 的精确度和 F1-score 低于 BFS-Max，社区密度低于 BM-adv，说明融合多种语义关系来定位社区有利于提高社区的质量。此外，与其他算法的社区密度值相比，BM-adv 的社区密度值提高较大，且 BM-adv 的精确度和 F1-score 仅次于 BFS-Max，牺牲一定的精确度换取更高的社区密度是有意义的。

4.2.2　社区搜索效率对比分析

本节将通过两种对比验证了 HCSMS 的效率：1)仅对比各种搜索算法定位社区的时间，不包含模型的训练时间，结果如图 4 所示；2)对比 HCSMS 使用 BFS-Max 算法和 ICS-GNN 使用 BFS-swap 算法在不同查询频率下的平均查询时间，结果如图 5 所示。

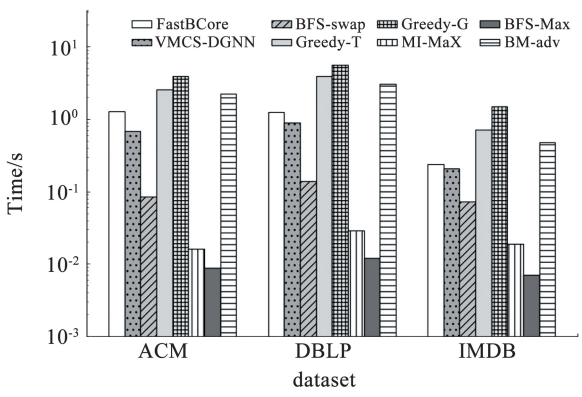


图 4　社区定位算法效率对比

Fig. 4　Comparison of efficiency of searching algorithms

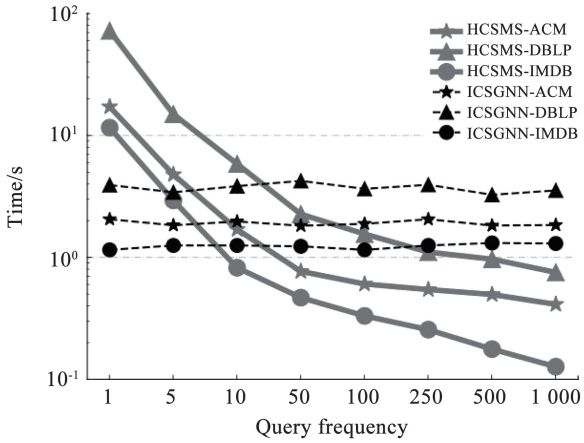


图 5 不同查询频率下的效率对比

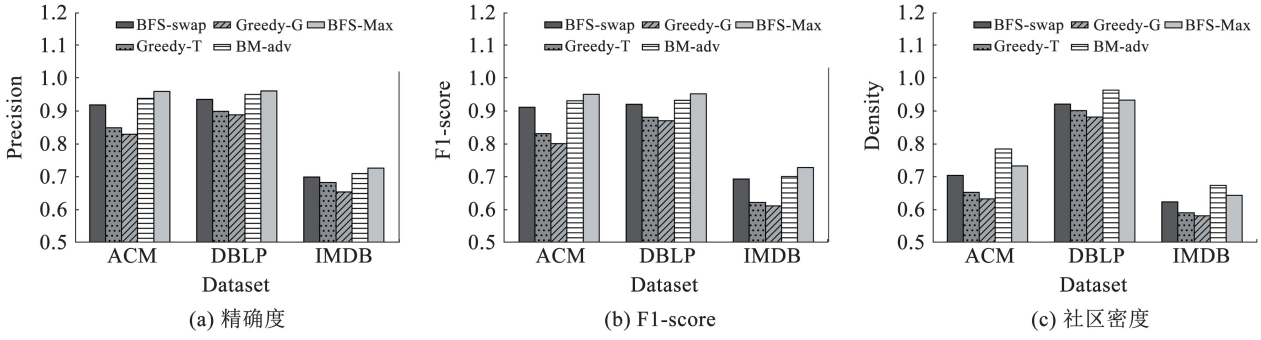
Fig. 5 Efficiency comparison under different query frequencies

从图 4 可以看到,BFS-Max 查询社区消耗的时间最少,其次是 MI-Max,再者是 BFS-swap。这是因为 BFS-swap 需要先捕获结构连通性,再交换节点,搜索时间比直接寻找最大评分和与最大互信息的 BFS-Max 和 BFS-MI 长。BFS-Max 由于通过网络优化减少了节点搜索范围,因此查询效率要比 BFS-MI 高。VMCS-DGNN 比 BFS-swap 耗时长,是因为 VMCS-DGNN 加入了 0/1 背包问题对节点价值进行计算。另外,BM-adv 的搜索时间也较长,这是因为 BM-adv 在定位最大评分和社区后,又进行了 k-core 判断。Greedy-T、Greedy-G 的搜索时间最长,因为其需要计算最短路径距离。

从图 5 可以看到,在各个数据集上 ICS-GNN 在不同查询频率下的平均查询时间几乎不变;在低频查询(1 ~ 10 次查询)时,ICS-GNN 的平均查询时间比 HCSMS 高,但随着查询频率增加,HCSMS 的平均查询时间逐渐减少。当查询频率达到 50 次时,HCSMS 的平均查询时间已经优于 ICS-GNN。这些结果体现了离线学习对于高频查询的优越性。

4.2.3 搜索算法分析

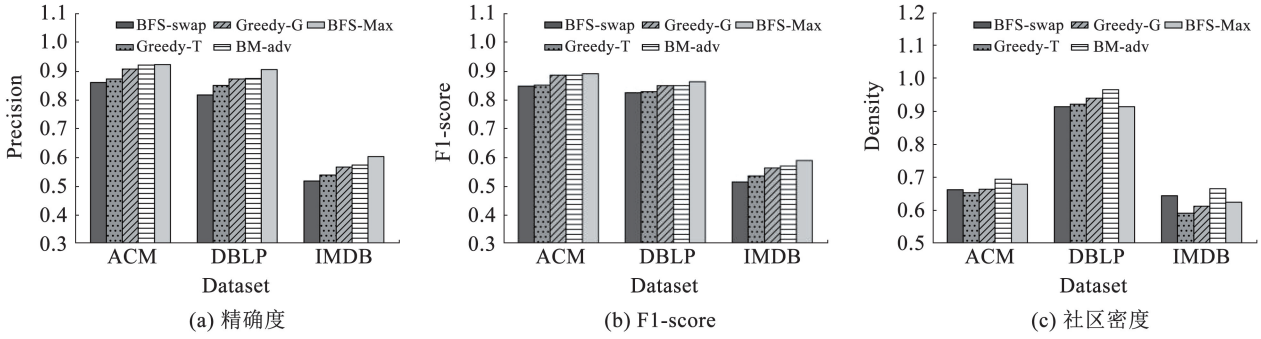
为了验证本文所提社区定位算法的有效性,本节在 HCSMS 中首先进行离线学习,然后分别用 BFS-Max、BM-adv、BFS-swap、Greedy-T、Greedy-G 定位社区,对比所获社区的质量,结果如图 6 所示;同理,在 ICS-GNN 中基于相同的 GNN 模型分别使用上述算法定位社区,对比所获社区的质量,结果如图 7所示。由图 6 和图 7 可以看到,BFS-Max 找到的社区具有最大的社区精确度和 F1-score,BM-adv 算法在社区密度上取得最大值,且 BM-adv 所获社区的精确度、F1-score 仅次于 BFS-Max,说明 BFS-Max 和 BM-adv 可以更好地判定社区成员。另外,对比图 6 和图 7 可以看到,同一个算法在 HCSMS 上的表现要比 ICS-GNN 更高,说明 HCSMS 的模型能够更好地捕获节点间的相似性。



(a) 精确度 (b) F1-score (c) 社区密度

图 6 HCSMS 模型上不同的社区定位算法对比

Fig. 6 Comparison of different community localization algorithms on the HCSMS model



(a) 精确度 (b) F1-score (c) 社区密度

图 7 ICS-GNN 模型上不同的社区定位算法对比

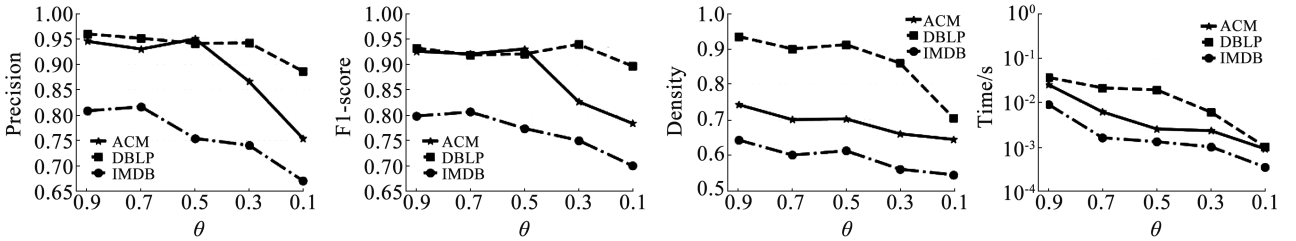
Fig. 7 Comparison of different community localization algorithms on the ICS-GNN model

4.3　参数敏感性分析

本文在 3 个数据集上测试了网络优化阈值 θ、社区规模 s、元路径数量 m 对社区精确度、F1-score 和社区密度以及社区搜索效率的影响。

4.3.1　网络优化阈值敏感性分析

网络优化阈值 θ 的取值范围在 0 ~ 1 之间，图 8 展示了不同 θ 下 BFS-Max 社区精确度、F1-score 和社区密度以及社区搜索效率。由图 8 可以看到，社区精确度、F1-score、社区密度和搜索时间均随 θ 的减小而减小，这是因为 θ 越小，网络优化时删除的连边会越多，网络越稀疏，尽管查询过程需要遍历的边减少，但真实的社区信息遭到破坏。在 ACM、DBLP 和 IMDB 数据集中，θ 分别取 0.5，0.3 和 0.7 左右可以同时兼顾社区质量和搜索时间。



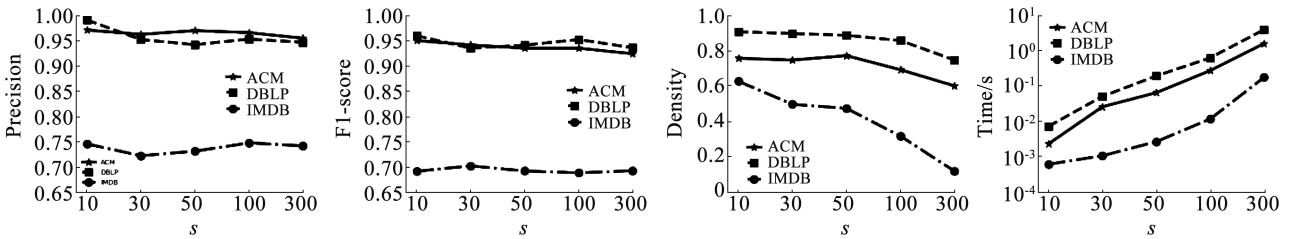
(a) 精确度　(b) F1-score　(c) 社区密度　(d) 效率

图 8　网络优化阈值分析

Fig. 8　Network optimization threshold analysis

4.3.2　社区规模敏感性分析

图 9 描述了 BM-Max 不同社区规模 s 下所获得的社区精确度、F1-score 和社区密度以及社区搜索效率。从图 9 可以看到，随着社区规模 s 的增加，3 个数据集上社区精确度和 F1-score 的变化幅度不大，但社区密度逐渐减小，查询过程需要的时间成本增加。



(a) 精确度　(b) F1-score　(c) 社区密度　(d) 效率

图 9　社区规模分析

Fig. 9　Community size analysis

4.3.3　元路径数量敏感性分析

表 3 给出了在不同 m 下 BFS-Max 所获得的社区精确度、F1-score 和社区密度以及社区搜索效率。其中 $m=1$ 表示在一条元路径下进行搜索。实验中社区规模 $s=100$。从表 3 可以看到，随着元路径数量的增加，社区精确度和 F1-score 以及社区密度都在增加，搜索时间也逐渐变长。这是因为元路径的增加扩展了节点间的交互关系，增加了网络的稠密性。

表 3　元路径数量敏感性分析

Tab. 3　Sensitivity analysis of the number of meta paths

数据集	MP	Precision	F1-score	Density	Time/s
ACM	$m=1$	0.91	0.91	0.652	0.214
	$m=2$	0.95	0.93	0.754	0.265
DBLP	$m=1$	0.83	0.80	0.705	0.364
	$m=2$	0.85	0.85	0.903	0.420
	$m=3$	0.95	0.96	0.962	0.607
IMDB	$m=1$	0.64	0.62	0.342	0.009
	$m=2$	0.71	0.68	0.673	0.014

4.4　消融研究

为了验证 HCSMS 各个模块的有效性，本文设计了 HCSMS 的 6 个变体进行消融实验，具体描述如下。

变体 1　$\text{HCSMS}_{\text{GNN}}$：HCSMS 中消除了 GNN 特征融合模块，仅考虑结构特征通过 BFS 挖掘规模为 s 的社区，以验证 HCSMS 通过 GNN 融合网络结构、节点属性信息查找社区的有效性。

变体 2　$\text{HCSMS}_{\text{sin}}$：HCSMS 中消除了多种语义关系影响，只使用一条元路径来寻找社区，以验证 HCSMS 融合多种语义关系查询社区的有效性。

变体 3　$\text{HCSMS}_{\text{node}}$：HCSMS 中消除了节点级别的关注，并赋予每个元路径邻居同样的重要性，以验证模型在节点级信息融合的有效性。

变体 4　$\text{HCSMS}_{\text{sim}}$：HCSMS 中消除了语义级别的关注，并赋予每条元路径同等的重要性，以验证模型区分多种语义关系重要性进行融合的有效性。

变体 5 $HCSMS_{re}$：HCSMS 中消除了网络优化模块的影响，在模型训练完成后直接通过 BFS-Max 定位目标社区，以验证网络优化对社区搜索效率的影响。

变体 6 $HCSMS_{BM}$：HCSMS 中消除了 BM-adv 算法对结构的限制，仅挖掘最大评分和社区。HCSMS 和各个变体的搜索结果如表 4 所示，可以看到，$HCSMS_{GNN}$的查询时间最短，但社区精确度最低，说明通过 GNN 捕获网络中多维信息综合衡量节点相关性对搜索高质量的社区是有效的。

HCSMS 的所有指标均比 $HCSMS_{sin}$ 高，且 $HCSMS_{sin}$的社区密度较其他变体都低，说明融合多种语义关系对社区搜索是有利的，且仅基于一种元路径难以挖掘到密度较高的社区。

HCSMS 的所有指标均比 $HCSMS_{node}$高，说明模型使用节点注意力融合有利于对节点做出更好的评分，对搜索高质量的社区是有效的。

HCSMS 的所有指标均比 $HCSMS_{sim}$高，说明区分多种语义关系的重要性来挖掘社区可以提高社区的精准度和结构凝聚性。

在所有变体中 $HCSMS_{re}$搜索耗时最长，说明网络优化可以降低查询的时间成本。

HCSMS 的社区密度是最高的，但精准度和 F1-score 仅次于 $HCSMS_{BM}$，说明挖掘结构凝聚性会影响到社区的属性相似性，但牺牲一定的精确度来换取高回报的社区凝聚性是有意义的。

表 4 消融研究结果

Tab. 4 Results of ablation research

数据集	指标	$HCSMS_{GNN}$	$HCSMS_{sin}$	$HCSMS_{node}$	$HCSMS_{sim}$	$HCSMS_{re}$	$HCSMS_{BM}$	HCSMS
ACM	Precision	0.560	0.900	0.910	0.930	0.950	**0.960**	0.940
	F1-score	0.390	0.890	0.900	0.920	0.930	**0.930**	0.930
	Density	0.720	0.650	0.710	0.700	0.730	0.670	**0.740**
	Time	**0.035**	0.020	0.100	0.100	0.250	0.090	0.120
DBLP	Precision	0.480	0.830	0.890	0.900	0.940	**0.950**	0.950
	F1-score	0.470	0.830	0.870	0.910	0.920	**0.920**	0.920
	Density	0.950	0.700	0.900	0.930	0.940	0.930	**0.960**
	Time	**0.080**	0.100	0.200	0.180	0.320	0.180	0.290
IMDB	Precision	0.500	0.620	0.560	0.630	0.710	**0.720**	0.720
	F1-score	0.500	0.600	0.500	0.520	0.600	**0.670**	0.640
	Density	0.630	0.340	0.590	0.600	0.650	0.640	**0.670**
	Time	**0.003**	0.009	0.010	0.010	0.032	0.011	0.012

4.5 案例研究

为了验证区分多种语义关系挖掘社区的有效性，本文选取 DBLP 数据集中的$\mathcal{P}_1 = APCPA$、$\mathcal{P}_2 = APTPA$ 两条元路径进行案例研究，并设置查询节点 $q =$“*WeiZhou*”进行了如下查询：1）分别基于元路径 $\mathcal{P}_1 = APCPA$ 和 $\mathcal{P}_2 = APTPA$ 单独查询；2）基于 $\mathcal{P}_3 = \mathcal{P}_1 + \mathcal{P}_2$ 查询，即同时考虑两种语义关系，但不区分两种语义关系的重要性；3）基于 $\mathcal{P}_4 = \varpi\mathcal{P}_1 + \sigma\mathcal{P}_2$（$\varpi + \sigma = 1, \varpi > \sigma$）查询，即同时考虑两种语义关系，但$\mathcal{P}_1$ 比$\mathcal{P}_2$ 重要；4）基于 $\mathcal{P}_5 = \sigma\mathcal{P}_1 + \varpi\mathcal{P}_2$（$\varpi + \sigma = 1, \varpi > \sigma$）查询，即同时考虑两种语义关系，但$\mathcal{P}_2$ 比$\mathcal{P}_1$ 重要。查询结果见表 5，可以看到，基于$\mathcal{P}_1$ 和 $\mathcal{P}_2$ 的查询均未找到满足社区规模 $s = 10$ 的社区，且基于$\mathcal{P}_1$ 和$\mathcal{P}_2$ 的查询结果中除“***Steve Jones***”和“***Zeng Minzu***”的重叠，其余节点都只出现在一条元路径的查询结果中；基于$\mathcal{P}_3$ 的查询可以找到满足规模的社区，且包含只在$\mathcal{P}_1$ 和$\mathcal{P}_2$ 独立查询结果中的作者；基于$\mathcal{P}_4$ 和$\mathcal{P}_5$ 的查询可以同时找到$\mathcal{P}_1$ 和$\mathcal{P}_2$ 独立查询结果中的作者，且$\mathcal{P}_4$ 的查询过程中优先选择只基于$\mathcal{P}_1$ 找到的作者，原因是$\mathcal{P}_4$ 的查询中$\mathcal{P}_1$ 比$\mathcal{P}_2$ 重要，反之，$\mathcal{P}_5$ 查询过程则优先选择只基于$\mathcal{P}_2$ 找到的作者。由于“***Steve Jones***”和“***Zeng Minzu***”是$\mathcal{P}_1$ 和$\mathcal{P}_2$ 独立查询结果中的重叠作者，表明他们与查询节点 $\boldsymbol{q}$ 的联系更密切，在$\mathcal{P}_4$ 和$\mathcal{P}_5$ 的查询中有较高的评分，会较早加入社区。

另外，为了验证 HCSMS 中离线学习和在线搜索策略的有效性，本文针对 3 个查询节点，分别进行了一次不分离学习和搜索的查询，即输入一个查询节点，根据其标签类别训练一个 GNN 模型获得节点评分来定位社区。然后，对 3 个查询节点分别再进行一次使用的离线学习和在线搜索策略的查询，即在离线阶段根据节点标签类别训练多个 GNN 模型，然后通过在线阶段输入查询节点，根据其标签类别

从学习好的 GNN 模型上获得节点评分来定位社区。由表 6 查询结果可见，不分离学习和搜索方式查找到的作者与分离学习和搜索查找到的作者是相同的，说明离线学习和在线搜索策略不会影响到社区成员的定位，因此，HCSMS 使用的离线学习和在线搜索策略是有效的。

表 5　语义关系案例研究

Tab. 5　Semantic Relationship Case Study

$\mathcal{P}_1 = APCPA$ ($s=10$)	$\mathcal{P}_2 = APTPA$ ($s=10$)	$\mathcal{P}_3 = \mathcal{P}_1 + \mathcal{P}_2$ ($s=10$)	$\mathcal{P}_4 = \varpi\mathcal{P}_1 + \sigma\mathcal{P}_2$ ($s=7$)	$\mathcal{P}_5 = \sigma\mathcal{P}_1 + \varpi\mathcal{P}_2$ ($s=7$)
Wei Zhou, *Xiaoya Tang*, ***Steve Jones***, *Arjan van Hessen*, *Yejun Wu*, *M. Cohen*, ***Zeng Minzu***, *WillemijnHeeren*	*Wei Zhou*, *Yan Qu*, ***Steve Jones***, *Po-Jun Tsai*, *Joshua Lewis*, ***Zeng Minzu***, *Huyen-Trang Vu*, Clinton Mah.	*Wei Zhou*, *Xiaoya Tang Yan Qu*, ***Steve Jones***, *Po-Jun Tsai*, *Arjan van Hessen*, ***Zeng Minzu***, *Joshua Lewis*, *Huyen-Trang Vu*, *WillemijnHeeren*	*Wei Zhou*, ***Steve Jones***, ***Zeng Minzu***, *XiaoyaTang*, *Arjan van Hessen*, *WillemijnHeeren*, *Po-Jun Tsai*	*Wei Zhou*, ***Steve Jones***, ***Zeng Minzu***, *Yan Qu*, *Po-Jun Tsai*, *Joshua Lewis*, *Huyen-Trang Vu*

表 6　搜索策略案例研究

Tab. 6　Semantic relationship case study

查询节点	*Herman Lam*	*Takeo Kanade*	*Alex Pentland*
不进行预训练查询	*Herman Lam*, *Benjamin G. Zorn*, *Linan Jiang*, *Malcolm P. Atkinson*, *Nathan Folkert*	*Takeo Kanade*, *Sanjeev Kumar*, *Ian N. Robinson*, *Roger Y. Tsai*, *J. David Schaffer*	*Alex Pentland*, *Sanjeev Kumar*, *Ian N. Robinson*, *Roger Y. Tsai*, *J. David Schaffer*
预训练查询	*Herman Lam*, *Benjamin G. Zorn*, *Linan Jiang*, *Malcolm P. Atkinson*, *Nathan Folkert*	*Takeo Kanade*, *Sanjeev Kumar*, *Ian N. Robinson*, *Roger Y. Tsai*, *J. David Schaffer*	*Alex Pentland*, *Sanjeev Kumar*, *Ian N. Robinson*, *Roger Y. Tsai*, *J. David Schaffer*

5　结　语

本文研究了异质信息网络中的社区搜索问题，提出了异质信息网络中融合多种语义关系的社区搜索方法 HCSMS，及采用离线学习和在线搜索的高效策略，从 HIN 中挖掘属性相似、结构内聚且包含多种语义关系的社区。HCSMS 通过多条元路径捕捉 HIN 的多种语义关系，并通过语义注意力机制区分各种语义关系的重要性来引导社区定位，离线阶段预训练节点-社区关联模型，生成节点归属各类社区的概率分布向量，在线查询则基于具体的查询，利用预训练模型快速定位目标社区。HCSMS 无需基于规则定义社区，具有较高的灵活性；其概率预计算机制与查询无关，显著提高了高频场景下的社区搜索效率。与 5 个基线算法对比及消融案例研究的结果，验证了 HCSMS 的有效性和效率。

目前，使用 GNN 进行异质网络社区搜索的研究尚处于探索阶段，异质网络中除结构信息、属性信息外，还包含许多信息，如时间信息、位置信息等，因此融合更多信息的异质网络社区搜索还有待深入研究。

参考文献

[1] FANG Yixiang, HUANG Xin, QIN Lu, et al. A survey of community search over big graphs[J]. The VLDB Journal, 2020, 29(1): 353

[2] 石川, 王睿嘉, 王啸. 异质信息网络分析与应用综述[J]. 软件学报, 2022, 33(2): 598
SHI Chuan, WANG Ruijia, WANG Xiao. Survey on heterogeneous information networks analysis and application[J]. Journal of Software, 2022, 33(2): 598. DOI: 10.13328/j.cnki.jos.006357

[3] 周丽华, 王家龙, 王丽珍, 等. 异质信息网络表征学习综述[J]. 计算机学报, 2022, 45(1): 160
ZHOU Lihua, WANG Jialong, WANG Lizhen, et al. A review of heterogeneous information network representation learning[J]. Chinese Journal of Computers, 2022, 45(1): 160. DOI: 10.11897/SP.J.1016.2022.00160

[4] FANG Yixiang, WANG Kai, LIN Xuemin, et al. Cohesive subgraph search over large heterogeneous information networks[M]. Springer, 2022: 96. DOI: 10.1007/978-3-030-97568-5

[5] FANG Yixiang, WANG Kai, LIN Xuemin, et al. Cohesive subgraph search over big heterogeneous information networks: Applications, challenges, and solutions[C]//Proceedings of the 2021 International Conference on Management of Data. Virtual Event, China 2021. New York, NY, USA: ACM, 2021: 2829. DOI: 10.1145/3448016.3457538

[6] FANG Yixiang, YANG Yi, ZHANG Wenjie, et al. Effective and efficient community search over large heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2020, 13(6): 854. DOI: 10.14778/3380750.3380756

[7] YANG Yixing, FANG Yixiang, LIN Xuemin, et al. Effective and efficient truss computation over large heterogeneous information networks[C]//2020 IEEE 36th International Conference on Data Engineering (ICDE). Dallas, TX, USA: IEEE, 2020: 85. DOI:

10.1109/ICDE48307.2020.00083

[8] DONG Zhen, HUANG Xing, YUAN Gang, et al. Butterfly-core community search over labeled graphs[J]. arXiv preprint arXiv: 2105.08628, 2021. DOI: 10.14778/3476249.3476258

[9] WANG Kai, ZHANG Wenjie, LIN Xuemin, et al. Efficient and effective community search on large-scale bipartite graphs[C]// 2021 IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece: IEEE. 2021. DOI: 10.1109/ICDE51399.2021.00015

[10] QIAO Lianpeng, ZHANG Zhiwei, YUAN Ye, et al. Keyword-centric community search over large heterogeneous information networks[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2021: 158. DOI: 10.1007/978-3-030-73194-6_12

[11] GAO Jun, CHEN Jiazun, LI Zhao, et al. ICS-GNN: Lightweight interactive community search via graph neural network[J]. Proceedings of the VLDB Endowment, 2021, 14(6): 1006. DOI: 10.14778/3447689.3447704

[12] WANG Xiao, JI Houye, SHI Chuan, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. Taipei, China: ACM, 2019: 2022. DOI: 10.1145/3308558.3313562

[13] HUANG Xin, CHENG Hong, QIN Lu, et al. Querying k-truss community in large and dynamic graphs[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, UT, USA: ACM, 2014:1311. DOI: 10.1145/2588555.2610495

[14] SOZIO M, GIONIS A. The Community-search problem and how to plan a successful cocktail party[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA: ACM, 2010: 939. DOI: 10.1145/1835804.1835923

[15] TENG Long, WANG Yanhao, LIN Zhe, et al. Topic-aware most influential community search in social networks[J]. Neuro Computing, 2025, 638: 130173. DOI:10.1016/j.neucom.2025.130173

[16] GUO Yaochen, GU Xiaoyan, WANG Zhuo, et al. RCS: An attributed community search approach based on representation learning[C]//2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE, 2021. DOI:10.1109/IJCNN52387.2021.9534285

[17] 赵卫绩, 张凤斌, 刘井莲. 一种基于节点嵌入表示学习的社区搜索算法[J]. 控制与决策, 2021, 36(8): 7
ZHAO Weiji, ZHANG Fengbin , LIU Jinglian. Community search algorithm based on node embedding representation learning[J]. Control and Decision, 2021, 36(8) :7. DOI:10.13195/j.kzyjc.2019.1439

[18] LIU Jinglian, WANG Daling, FENG Shi, et al. Learning distributed representations for community search using node embedding[J]. Frontiers of Computer Science, 2019, 13(2): 3

[19] CHEN Jie, GAO Jun, CUI Bin. ICS-GNN +: Lightweight interactive community search via graph neural network[J]. The VLDB Journal, 2023, 32(2): 447. DOI: 10.1007/s00778-022-00754-0

[20] JIANG Yu, RONG Yu, CHENG Hong, et al. Query driven-graph neural networks for community search: From non-attributed, attributed, to interactive attributed[J]. arXiv preprint: 2104.03583, 2021. DOI: 10.14778/3514061.3514070

[21] 王亚峰, 周丽华, 陈伟, 等. 异质信息网络的互信息最大化社区搜索[J]. 浙江大学学报(工学版), 2023, 57(2): 287
WANG Yafeng, ZHOU Lihua, CHENG Wei, et al. Community search with mutual information maximization over heterogeneous information networks[J]. Journal of Zhejiang University (Engineering Science), 2023, 57(2): 287. DOI: 10.3785/j.issn.1008-973X.2023.02.009

[22] 陈伟, 周丽华, 王亚峰, 等. 异质信息网络中基于解耦图神经网络的社区搜索[J]. 计算机科学, 2024, 51(3): 90
CHEN Wei, ZHOU Lihua, WANG Yafeng, et al. Community search based on decoupled graph neural networks in heterogeneous information networks[J]. Computer Science, 2024, 51(3): 90. DOI: 10.11896/jsjkx.221200029

[23] LI Yuqi, ZANG Guo sheng, SONG Chunyao, et al. Leveraging semantic information for enhanced community search in heterogeneous graphs[J]. Data Science & Engineering, 2024, 9(3): 220. DOI: 10.1007/s41019-024-00244-z

[24] SONG Yixin, ZHOU Lihua, YANG Peizhong, et al. CS-DAHIN: Community search over dynamic attribute heterogeneous network[J]. IEEE Transactions on Knowledge and Date Engineering, 2024, 36(11): 15. DOI: 10.1109/TKDE.2024.3402258

[25] WANG Xiao, LIU Nian, HAN Hui, et al. Self-supervised heterogeneous graph neural network with co-contrastive learning[C]// Proceedings of the 27th ACM SIGKDD. International Conference on Knowledge Discovery and Date Mining. Singapore: ACM, 2021: 1726. DOI: 10.1145/3447548.3467415

[26] 竺俊超, 王朝坤. 复杂条件下的社区搜索方法[J]. 软件学报, 2019, 30(3): 21. DOI: 10.13328/j.cnki.jos.005699
ZHU Junchao, WANG Chaokun. Approaches to community search under complex conditions[J]. Journal of Software, 2019, 30(3): 21. DOI: 10.13328/j.cnki.jos.005699

[27] WANG Jialong, ZHOU Lihua, WANG Xiaoyu, et al. Attribute-sensitive community search over attributed heterogeneous information networks[J]. Expert Systems with Applications, 2024, 235: 121153. DOI:10.1016/j.eswa.2023.121153

(编辑 吕雪梅)