属性网络中结合用户偏好的社区搜索和离群点检测

李青青[1],马慧芳[1,2],李 举[1],李志欣[3],姜彦斌[1]
(1. 西北师范大学计算机科学与工程学院,甘肃兰州 730070; 2. 桂林电子科技大学广西可信软件重点实验室,广西桂林 541004;
3. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004)

摘 要: 社区搜索是备受关注的网络分析任务之一,旨在搜寻包含查询节点的局部社区.现有大多数社区搜索方法多面向简单网络且仅能定位查询节点所在社区,未能在搜索过程中考虑用户偏好.为实现利用用户偏好指导搜索过程并搜寻用户感兴趣的多社区,设计了属性网络中结合用户偏好的社区搜索和离群点检测方法,旨在通过较少的查询节点有效的捕获用户偏好并自动探索网络中的社区,同时识别社区中离群点.具体而言,通过编码查询节点及其邻居间的显式交互关系和相似属性以突出局部结构,利用其来挖掘潜在查询节点候选集成员.在查询节点候选集上定义平均划分相似度以推断属性子空间为用户潜在兴趣.采用属性和结构约束来搜索网络中的多社区和离群点.此外,真实数据集和人工数据集上的大量实验证明了所提方法的有效性.

关键词: 属性网络;社区搜索;平均划分相似度;属性子空间;离群点
中图分类号: TP181;TP274 **文献标识码:** A **文章编号:** 0372-2112(2022)09-2172-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210370

Incorporating User Preference for Community Search and Outlier Detection in Attributed Network

LI Qing-qing[1], MA Hui-fang[1,2], LI Ju[1], LI Zhi-xin[3], JIANG Yan-bin[1]
(1. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*;
2. *Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Science and Technology, Guilin, Guangxi* 541004, *China*;
3. *Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*)

Abstract: Community search aims to search local communities containing query nodes, which is one of the most concerned studies in network analysis task. Most existing community search methods are oriented to simple network and can only detect the community where query nodes are located. They may fail to take user's preferences into account during searching process. To guide the process of community search via user's preferences for finding multi-communities that users are interested in, we propose a community search method that is capable of searching multi-communities with user's preference and simultaneously identify outliers via few given query nodes in attributed network. Clearly, we explicitly model interactions between query nodes with its neighbors and encode similar attributes to highlight the local structure, which could be beneficial for query nodes to mine potential candidates. And we define the average partition similarity on candidate set of query nodes to infer attribute subspace as user's latent interest. Multi-communities and outliers in the whole network are detected via fractional-core and structural constraints. Experiments on real and synthetic network datasets demonstrate the effectiveness of the proposed algorithm.

Key words: attributed network; community search; average partition similarity; attribute subspace; outliers

1 引言

作为分析复杂系统本质和功能的有力表示,图可以建模各种复杂系统单元之间的交互关系.其中节点表示复杂系统单元,边表示节点间潜在结构和交互规则.此外,节点往往可以与大量属性相关联,描述节点的特定特征并提供与网络拓扑正交的有价值的信

收稿日期:2021-03-18;修回日期:2021-11-13;责任编辑:梅志强
基金项目:国家自然科学基金(No.61762078,No.61363058,No.6196604);广西多源信息挖掘与安全重点实验室开放基金(No.MIMS18-08);西北师范大学2019年度青年教师科研能力提升计划重大项目(No.NWNU-LKQN2019-2)

息[1,2]. 作为网络分析中的一个基本问题,社区搜索旨在挖掘与用户给定查询节点相关联的局部社区[3]. 与社区检测[4]不同,社区搜索将个性化需求整合到社区搜索过程中在特定用户挖掘[5]、蛋白分析网络[6]等任务中具有广泛的应用前景.

传统的社区搜索方法主要集中于普通图(无属性)[7~9]. 随着现实世界属性信息的激增,属性可作为辅助信息潜在的提高社区搜索的准确性. 早期的属性社区搜索方法将属性视为同等重要,旨在通过属性相似性找到包含查询节点的局部内聚社区. 如 Leman 等人[10]提出的基于属性图的无参数化方法,即 PICS(Parameter-free Identification of Cohesive Subgroups),将所有属性视为节点来挖掘局部社区. Shang 等人[11]设计了既能反映网络拓扑结构关系也能包含属性相似性的 TA-graph 来检测局部社区. 但高维的属性信息使得这类方法的存储开销加剧. 因此,有研究人员指出查询节点所携带的核心属性足以指导算法挖掘出融合用户偏好的个性化局部社区[12]. 在不影响精确性的前提下,可以使用核心属性进行个性化属性社区搜索. 如 Huang 等人[13]研究了满足结构内聚性 k-core 和关键字内聚性的属性社区. VAC(Vertex-centric Attributed Community)[14]旨在挖掘包含查询节点且具有最大属性得分的连通 k-truss. Ye 等人[15]通过引入统计学方法来搜索具有指定属性的查询节点所在社区. 同时,目标社区发现方法也因其能捕获个性化需求而被广泛研究. 如 Perozzi 等人[16]提出了面向用户的属性图挖掘方法,利用焦点属性以提取目标社区与离群点. TSCM(Target Subspace and Communities Mining)[17]通过属性子空间来定位和挖掘目标社区. 尽管以上方法缓解了存储开销,但大多数方法均仅能定位查询节点所在社区,未整合用户偏好到社区搜索过程以挖掘融合用户偏好的多社区且精准定位与社区中成员紧密连接但属性偏离其社区成员的节点. 此外,尽管部分方法考虑了整合用户偏好到社区搜索过程中,但仍需用户提供足够多的查询节点来帮助算法捕获用户偏好,具有灵活性不足且不切合实际的局限性.

针对以上问题,提出了属性网络中结合用户偏好的社区搜索和离群点检测方法(Incorporating user Preference for Community Search and Outlier detection in attributed network, IPCSO). 具体地,通过编码查询节点邻域网络中节点属性和结构间关系来捕获潜在社区成员. 其次,定义平均划分相似度来挖掘属性子空间,并将其作为用户偏好来指导社区搜索过程. 最后,将属性子空间蕴含的重要信息融入到网络中,并采用结构凝聚力约束 k-core 和属性内聚性约束 fractional-core 来搜索网络中的多社区并检测离群点. 多种真实网络和人工网络上的广泛实验证明了本文方法的有效性和效率.

2　准备知识

2.1　符号说明与问题定义

给定无向加权属性图 $G=(V,E,T,\boldsymbol{W})$,其中 $V=\{v_i\}_{i=1,\cdots,n}$ 表示图中节点集且 $|V|=n$;$E\subseteq V\times V$ 表示边集且 $|E|=m$. 属性集为 $T=\{t_i\}_{i=1,\cdots,d}$,$|T|=d$. 设节点属性矩阵为 $\boldsymbol{F}\in\mathbb{R}^{n\times d}$,$\boldsymbol{f}_i^{\mathrm{T}}$ 表示节点 v_i 的属性向量. 权重矩阵记为 $\boldsymbol{W}\in\mathbb{R}^{n\times n}$,其中 $w_{ij}=\cos(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 表示边 (v_i,v_j) 的权重. 设用户给定的查询节点集为 $Q=\{q_i\}_{i=1,\cdots,s}$,$Q\subseteq V$. 属性约束 fractional-core 的阈值为 w,结构约束 k-core 的阈值为 k. IPCSO 的目标是找到目标社区集 $C=\{C_i\}_{i=1,\cdots,l}$ 和离群点集 $O=\{O_i\}_{i=1,\cdots,b}$,满足:(1)社区内紧密连接;(2)在属性上与用户偏好(属性子空间)一致;(3)找到偏离属性约束的社区内成员(离群点).

具体地,本文方法的问题定义如下:(1)输入:属性图 $G=(V,E,T,\boldsymbol{W})$,用户提供的查询节点集合 Q,属性约束阈值 w 以及结构约束阈值 k. (2)输出:融合用户偏好的社区集 C 与离群点集 O.

2.2　结合用户偏好的社区搜索和离群点检测基本框架

本文提出的方法框架如图 1 所示:在给定属性网络 $G=(V,E,T,\boldsymbol{W})$,查询节点集 Q,结构约束阈值 k 和属性约束阈值 w 的情况下,采用编码可以显式建模邻居之间的交互以突出局部结构内的公共属性,有助于算法挖掘潜在社区成员. 通过平均划分相似度获取每个属性的重要性权重,以此表示用户偏好. 通过属性子空间的指导对网络重加权,并设定结构约束及属性约束检测多社区以及社区中的离群点. 接下来将详细介绍该算法.

定义 1(节点 v_i 的邻域网络)　给定节点 v_i,其邻域网络被定义为 $N(v_i)=(V_N(v_i),E_N(v_i),T_N(v_i),\boldsymbol{W}_N(v_i))$,其中节点集为 $V_N(v_i)=\{v_w|(v_i,v_w)\in E\}\cup\{v_i\}$,$E_N(v_i)=\{(v_u,v_w)|v_u\in V_N(v_i)\wedge v_w\in V_N(v_i)\wedge(v_u,v_w)\in E\}$ 为边集,$T_N(v_i)$ 表示 $V_N(v_i)$ 的属性集,$T_N(v_i)\subseteq T$. $\boldsymbol{W}_N(v_i)\in\mathbb{R}^{|V_N(v_i)|\times|V_N(v_i)|}$ 为 $E_N(v_i)$ 的权重矩阵.

3　结合用户偏好的社区搜索和离群点检测

3.1　编码节点属性和结构

较少的查询节点(例如,一个查询节点)包含的信息有限,无法准确计算属性子空间. 此外,查询节点间可能没有相互作用关系,无法保证推断出的子空间与内聚连接相关联. 针对此,受 CDE 模型[18]的启发,设计了建模查询节点与其邻居间相互作用及属性相关性的方法. 该方法可有效地挖掘查询节点所在社区的潜在

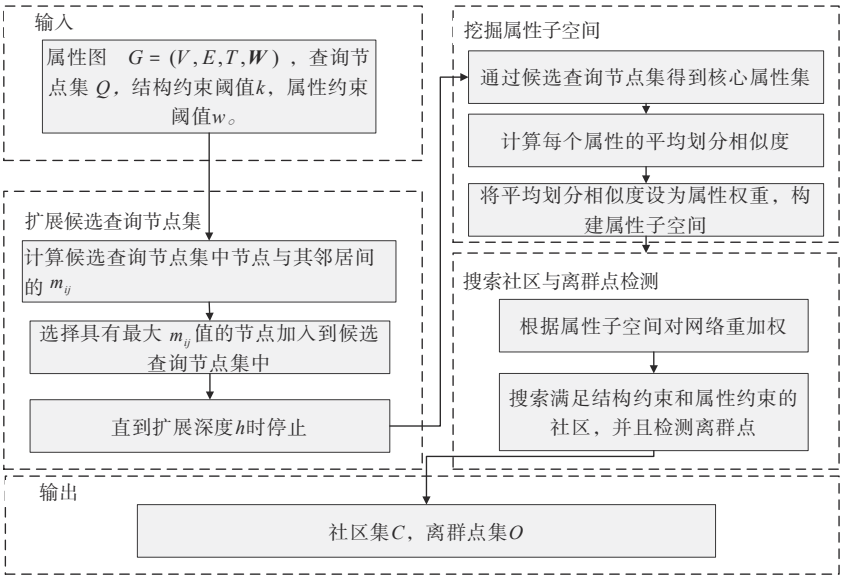


图1 结合用户偏好的社区搜索和离群点检测的基本框架

成员,将较少的查询节点扩展为一组候选查询节点. 具体地,编码节点结构和属性信息以增强节点的表示能力,从而避免丢失与查询节点相似的潜在社区成员信息. 因此,尽管给定少量查询节点,查询节点的特征信息仍可保留用以查找与其相似的候选节点.

接下来将详细介绍该策略. 首先,编码节点结构和属性相似性m_{ij}的公式被定义为:

$$m_{ij}=\max\{\ln(w_{ij}D/(d_id_j))-\ln a,0\} \tag{1}$$

其中$d_i=\sum_{j\in V_N(v_i)}w_{ij}$,表示节点$v_i$与其邻域网络中所有节点之间的相似度之和. $D=\sum_i d_i$是邻域网络中所有节点的相似度之和. 本文设a=2.

具体地,首先提取节点v_i的邻域网络. 然后,利用式(1)对每个候选查询节点及其邻居节点的结构和属性关系进行编码. 最后,将m_{ij}值最大的节点v_j加入候选节点集中. 上述过程持续进行到候选节点的数量大于扩展深度h为止. 图2显示了编码节点属性和结构的过程,其中边的粗度表示端点节点的属性相似度. 设h=5,v_2为用户提供的查询节点,首先初始化候选查询节点集$Q_1=\{v_2\}$并提取v_2的邻域网络. 其次,通过式(1)对查询节点与其邻居节点间的相似度进行编码,选择具有最大m_{2j}的节点v_j添加到候选查询节点集中,得$Q_1=\{v_2,v_3\}$. 重复上述步骤,分别提取v_2和v_3的邻域网络,计算v_2与其邻居$v_j\in V_N(v_2)$间的相似强度m_{2j},v_3与其邻居$v_a\in V_N(v_3)$间的相似强度m_{3a},选择节点v_8加入到候选查询节点集中. 上述过程在候选查询节点个数大于5时停止,最后得候选查询节点集为$Q_1=\{v_1,v_2,v_3,v_4,v_8\}$.

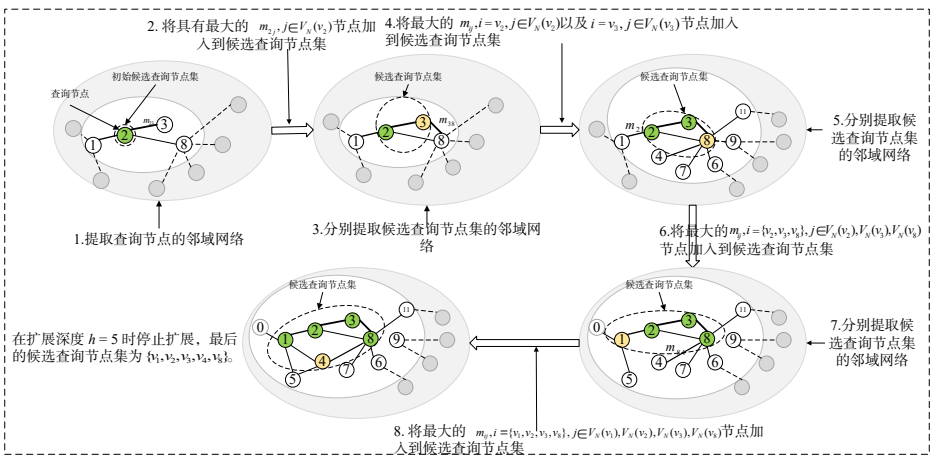


图2 编码节点属性和结构策略的示例

3.2 基于平均划分相似度挖掘属性子空间

本节设计了一种属性子空间推断方法,该方法适用于衡量真实社区中属性的重要性. 首先通过扩展后的查询节点集来提取用户所关注的核心属性,随后根据核心属性集挖掘用户所聚焦的属性子空间,以此来指导算法挖掘融合用户偏好的社区. 特别地,基于距离属性划分方法[19,20]的潜在含义,同一社区中的属性应产生相似的划分,存在一些由属性定义的划分与实际节点所属社区一致. 首先根据候选查询节点集将其中节点所携带的属性提取为核心属性集 T_{core}, $T_{\text{core}}\subseteq T$. 对于 $\forall t_i\in T_{\text{core}}$,由属性 t_i 划分的节点集 V 表示为 $V/t_i=\{C_i, \Gamma_i\}$, C_i 表示包含属性 t_i 的节点集, Γ_i 表示不包含属性 t_i 的节点集,且 $C_i\cup\Gamma_i=V$.

定义2(划分熵) 给定属性 $t_i\in T_{\text{core}}$,设属性 t_i 的划分为 $V/t_i=\{C_i, \Gamma_i\}$. $P(C_i)$ 表示包含属性 t_i 的节点集合的概率,则属性 t_i 的划分熵 $E(t_i)$ 定义为:

$$E(t_i)=-P(C_i)\log_2 P(C_i)-P(\Gamma_i)\log_2 P(\Gamma_i) \tag{2}$$

定义3(划分条件熵) 给定核心属性 t_i, $t_j\in T_{\text{core}}$,设属性 t_i 和 t_j 的划分分别为 $V/t_i=\{C_i, \Gamma_i\}$, $V/t_j=\{C_j, \Gamma_j\}$. 则核心属性 t_j 关于属性 t_i 的划分条件熵 $\text{CE}_{t_i}(t_j)$ 定义为:

$$\text{CE}_{t_i}(t_j)=-\sum_{x\in V/t_j}P(x)\sum_{y\in V/t_i}P(y|x)\log_2 P(y|x) \tag{3}$$

定义4(划分联合熵) 给定核心属性 $t_i, t_j\in T_{\text{core}}$,核心属性 t_i 和 t_j 的划分联合熵 $E(t_i, t_j)$ 定义为:

$$E(t_i,t_j)=E(t_j)+\text{CE}_{t_j}(t_i) \tag{4}$$

定义5(属性划分距离) 划分 V/t_i 和 V/t_j 的属性划分距离 $d(V/t_i, V/t_j)$ 为:

$$d(V/t_i,V/t_j)=\text{CE}_{t_i}(t_j)+\text{CE}_{t_j}(t_i) \tag{5}$$

定义6(归一化属性划分相似度) 为了便于度量属性划分的相似性,对属性划分距离进行归一化并将其转化为属性划分相似度:

$$S(V/t_i,V/t_j)=1-\frac{d(V/t_i,V/t_j)}{E(t_i,t_j)} \tag{6}$$

值得注意的是,式(6)的分母不能为零,所以不会出现未定义的情况. 此外,由于式(6)的分母总是大于分子,因此不会因为分母过大而放小比值.

由上可知,平均划分相似度可形式化为:

定义7(平均划分相似度) 给定属性 t_i, $t_j\in T_{\text{core}}$,属性 t_i 的平均划分相似度定义为:

$$\text{APS}(t_i)=\frac{\sum_{j=1,j\neq i}^{|T_{\text{core}}|}S(V/t_i,V/t_j)}{|T_{\text{core}}|-1} \tag{7}$$

APS衡量划分相似度差异性,可表示选用某一属性 t_i 进行划分与其他属性进行划分的相异程度. 属性 t_i 的APS值越大,则表明选用属性 t_i 划分所得的社区与其他属性划分所得社区的相似程度越高. 用向量 $\boldsymbol{\tau}$ 来表示属性子空间,其元素值计算如下:

$$\tau_i=\begin{cases}\text{APS}(t_i), & t_i\in T_{\text{core}}\\ 0, & t_i\notin T_{\text{core}}\end{cases} \tag{8}$$

元素 τ_i 表示对应属性 t_i 在属性子空间中的重要性或与查询节点的相关性,若属性 $t_i\in T_{\text{core}}$,则属性子空间中元素设为 $\text{APS}(t_i)$,否则,将其在属性子空间下的重要性设为0.

3.3 编码节点属性和结构

3.3.1 网络重加权

属性子空间可帮助用户探索在核心属性上内聚的社区,通过重加权网络将属性子空间对属性的关注程度融入到整个网络中. 首先定义在属性子空间的指导下重加权后的权重矩阵 $\boldsymbol{W}^{\boldsymbol{\tau}}$, $\boldsymbol{W}^{\boldsymbol{\tau}}$ 中元素 $w_{ij}^{\boldsymbol{\tau}}$ 表示节点 v_i 和 v_j 在属性子空间下的相似度. 具体计算公式如下:

$$w_{ij}^{\boldsymbol{\tau}}=\begin{cases}\text{sim}_{\boldsymbol{\tau}}(v_i,v_j), & (v_i,v_j)\in E\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

其中 $\text{sim}_{\boldsymbol{\tau}}(v_i,v_j)=\text{e}^{-\sqrt{(f_i-f_j)^{\text{T}}\,\text{diag}(\boldsymbol{\tau})(f_i-f_j)}}$,指数部分表示属性子空间 $\boldsymbol{\tau}$ 下的加权欧几里得距离.

3.3.2 社区搜索与离群点检测

现有方法常使用 k-core[21], k-truss[22] 和 k-clique[23] 等结构约束来度量社区结构的内聚性. k-core考虑了图中节点的结构内聚性,fractional-core[24]不仅考虑了图中节点的度,而且也考虑节点间的连接强度. 因此,选用前者来度量社区的结构凝聚力,后者度量属性子空间下端点节点属性的凝聚性. 具体定义如下:

定义8(k-core[21]) 给定一个正整数 k($k\geq 0$),图 G 的 k-core 表示为 H_k,其是图 G 的最大子图,使得 $\forall v\in H_k, \deg_{H_k}(v)\geq k$,其中 $\deg_{H_k}(v)=|V_N(v)-v|$.

值得注意的是 H_k 可能不是一个连通子图,设 k-core 的连通分量为 $\widehat{k\text{-core}}$.

定义9(fractional-core[24]) 给定一个有理数 w, fractional-core是图 G 的最大子图 H_w, H_w 中节点的度都不小于 w,即 $\forall v_i\in H_w, d_i\geq w$,其中 $d_i=\sum_{j\in V_N(v_i)}w_{ij}^{\tau}$.

IPCSO方法旨在通过属性子空间的指导搜索出满足fractional-core和 k-core约束的多社区,同时识别出社区中的离群点,且所挖掘到的社区都是最大的. 也就是说,不存在另一个满足结构和属性内聚性约束的社区 $C\subset C'$. 具体地,首先提取满足结构约束 k-core 的所有密集连通子图作为候选社区. 其次,在这些密集连通子图上设置属性约束以满足属性内聚性. 最后,返回满足结构和属性约束的社区,其中不满足属性约束的节点为离群点.

4 实验结果与分析

本节将通过实验验证IPCSO方法的有效性和效率,旨在回答以下三个研究问题. 问题1:参数的变化对于IPCSO方法影响如何? 问题2:IPCSO方法相比其他基准方法的性能表现如何? 以及问题3:IPCSO方法在实际应用中表现如何?

4.1 数据集描述

4.1.1 人工数据集

为研究IPCSO方法的性能,基于LFR[25]生成了节点数为n、属性数为d的具有真实基准的人工网络. 其中人工网络中节点的平均度数由d_{avg}来控制,最大度通过d_{max}来设定. 此外,μ控制社区内边数与社区间边数的比值. 社区的大小分别通过参数c_{max}和c_{min}来控制. 给定社区中所需节点数,邻接矩阵对角线上定义的块以0.35的概率随机为块中的每个元素分配一条边. 对于非对角线上的块,以0.01的概率随机分配边. 更进一步地,按照均值为[0,1]以及方差控制在0.001范围内的高斯分布来分配属性值. 较小的方差便于社区中的节点包含有核心属性. 其余属性可从方差较大的正态分布中提取. 离群点可从每个社区中随机选择一个节点,将其属性替换为与该社区属性差异较大的属性. 经过多次测试,人工网络参数设置为:d_{avg}=20,d_{max}=80,c_{min}=2,c_{max}=80. 具体设置如表1所示.

表1 人工数据集统计信息

数据集	节点	边	μ	属性
Synth 1	1500	34950	0.1	$\{f_1,f_2,f_3,f_4,f_5,f_6\}$
Synth 2	8000	11853	0.15	$\{f_7,f_8,f_9\}$

4.1.2 真实数据集

选取了5个真实属性网络. 4Area是计算机科学领域的合著网络,其中节点表示作者,边表示作者间的合著关系,属性表征作者在数据库、信息检索和机器学习等方面发表的会议. YouTube是一个社交网络的视频分享网站,节点表示用户,边表示友谊关系,属性描述用户的身份信息. 而Disney数据集是共同购买网络,其中节点为影片,边表示共同购买关系,每部电影有28个属性. 由联邦能源管理委员会发布的Enron中,节点表示电子邮件地址,电子邮件之间的传输关系记为边. 节点上携带18个属性,用于描述邮件的平均内容长度、平均收件人数量等信息. 此外,IMDB(Internet Movie DateBase)数据集中节点表示电影,边表示电影中演员的共同参演关系. 具体统计数据如表2所示.

表2 真实数据集统计信息

数据集	节点数	边数	属性数	社区数
4Area	26144	108550	4	50
YouTube	77381	367151	30087	8385
Disney	124	333	28	9
Enron	13533	176967	18	40
IMDB	862	4388	21	30

4.2 实验设置

4.2.1 评价指标

为评估IPCSO方法在属性图上的有效性,采用Precision和CAS(Community Attribute Similarity)[14]来度量社区的质量. 设算法搜索到的社区为C,基准社区为C',则Precision=$|C'\cap C|/|C|$,其他评价指标定义如下:

定义10(CAS) CAS度量社区C中的属性凝聚力,形式上:

$$\mathrm{CAS}(C)=\frac{1}{|V_C|^2}\sum_{v_i\in V_C}\sum_{v_j\in V_C}\frac{|T(v_i)\cap T(v_j)|}{|T(v_i)\cup T(v_j)|} \tag{10}$$

其中$T(v_i)$为节点v_i所携带的属性集合. CAS值越高,属性内聚性越强.

4.2.2 对比方法

为度量IPCSO方法的性能,选取了以下三类方法进行对比. 第一,比较本文方法与目标社区发现方法的性能,选取了FocusCO(Focused Clustering and Outlier)和TSCM. FocusCO为经典的目标社区发现算法之一,TSCM与IPCSO均采用了扩展查询节点的策略来挖掘社区. 第二,为研究IPCSO与其他基于结构度量的属性社区搜索方法的性能,选择了ACQ(Attributed Community Query)和VAC-Etruss(Vertex-centric Attributed Community-Etruss). 其中ACQ是具有奠基性的社区搜索方法之一,而VAC-Ecore(Vertex-centric Attributed Community-Ecore)是以节点为中心的属性社区搜索方法的变体. 第三,为了探索IPCSO方法挖掘融合用户偏好社区的有效性,选取了LOCLU(LOcal CLustering Unimodality)方法.

4.3 参数分析(问题1)

本节将从三方面研究不同参数设置对于IPCSO的影响. 探索查询节点个数对于IPCSO的性能影响. 研究编码查询节点结构和属性策略中超参数a的选定以及扩展深度h的选择对于IPCSO的影响. 观察k和w的变化对于IPCSO性能的影响.

对于查询节点个数s,从每个真实数据集中任意选取6个查询节点,在s取值不同的情况下运行方法50次并取其平均值返回结果. 如图3所示. 不同数据集上查询节点个数的变化对于本文方法的影响均较小. 原因在于IPCSO方法使用较少的查询节点就可以有效的编码节点信息以找到潜在社区成员,有助于增强用户偏好以达到较为精确地搜索结果. 此外,由于所查找社区的属性内聚性通过用户指定的属性约束来控制,故查询节点个数的变化对于CAS的影响较小.

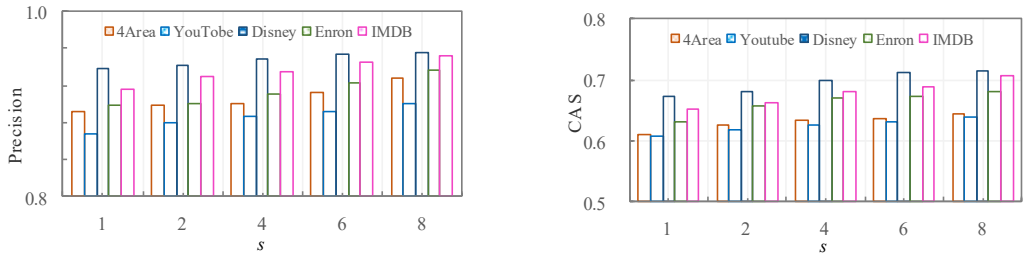
图4展示了编码查询节点结构和属性策略中的超

参数 a 和扩展深度 h 取值的影响. 从图 4(a)和图 4(b)均可观察到,a 和 h 的增加首先会带来较大的性能提升,但随着其取值不断增大会使得 IPCSO 的性能下降. a=2 时所有数据集上的性能表现较好,当 a=6 时部分数据集上的性能下降. 这说明当 a 设定为 2 时编码节点属性和结构策略足以帮助 IPCSO 找到候选查询节点集以提取用户偏好. 当 h=6 时,较小的数据集上的性能达到了最优. 在较大数据集中,h=8 时性能表现最好,之后则出现了性能下降. 究其原因在于扩展过深会使得将不属于查询节点所在社区的节点纳入.

图 5 给出了结构约束阈值 k 和属性约束阈值 w 分别从 8 以 4 的步长变化到 20,0.2 以 0.2 的步长变化到 0.8 的结果. 图 5(a)中可观察到,k 值的增加会使得算法所找到社区的结构凝聚性增强,其精度也会随之增加. k 的大小与返回社区的大小密切相关. 图 5(b)可观察到随着 w 的增加,算法的精度随之降低,源于较小的属性阈值将会使得所返回社区的属性内聚性更强,精度更高. 但如果将 w 设置过小,算法将会返回一个空社区. 本文设置 k=8,w=0.7.

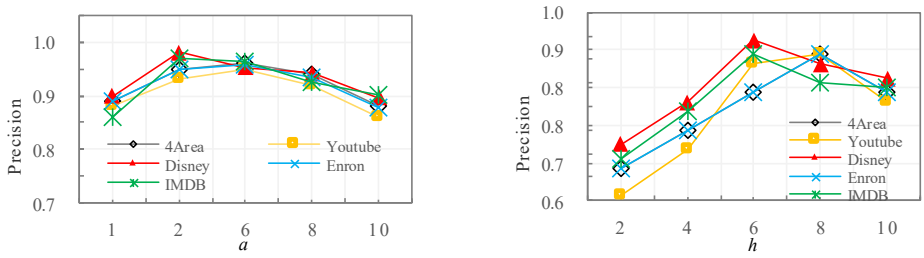
4.4　性能比较(问题 2)

本节将评价 IPCSO 方法的有效性,观察 IPCSO 方法与基准方法的性能差异.



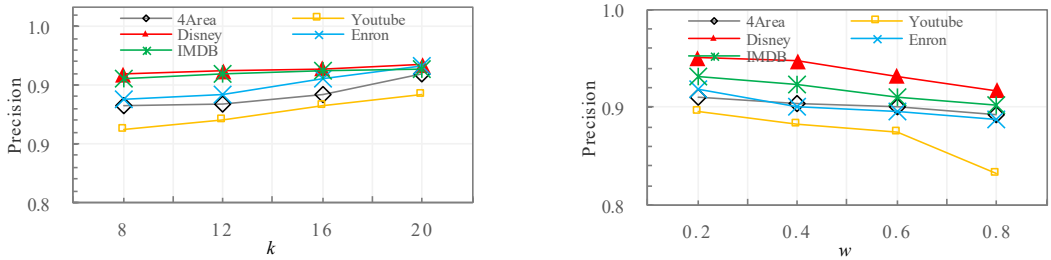
(a) Precision　　(b) CAS

图 3　不同数据集上 s 取值对应的 Precision 和 CAS



(a) a 取值对应的 Precision　　(b) h 取值对应的 Precision

图 4　不同数据集上 a 和 h 取值对应的 Precision



(a) k 取值对应的 Precision　　(b) w 取值对应的 Precision

图 5　不同数据集上 k 和 w 取值对应的 Precision

表 3 给出了在不同数据集上取 4 个查询节点,运行 100 次的平均结果. 从实验结果可得,TSCM 在所有数据集上的表现都优于 FocusCO. 因为较少的查询节点不足以精确地捕获到用户偏好,进而使得算法不能准确

表3 IPCSO与基线方法的整体性能比较

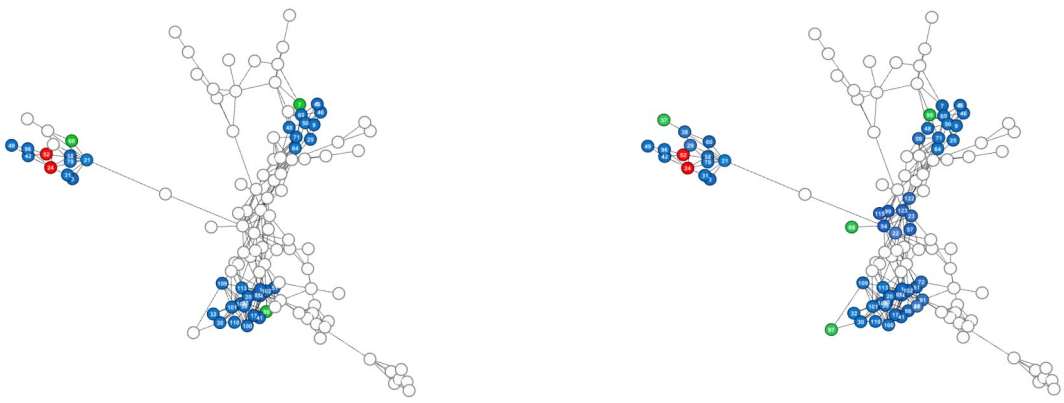
方法	评价指标	4Area	YouTube	Disney	Enron	IMDB	Synth 1	Synth 2
FocusCO	Precision	0.687	0.662	0.749	0.692	0.726	0.738	0.712
	CAS	0.532	0.514	0.598	0.543	0.571	0.581	0.558
TSCM	Precision	0.721	0.699	0.793	0.742	0.783	0.788	0.759
	CAS	0.546	0.531	0.593	0.556	0.582	0.582	0.567
ACQ	Precision	0.799	0.779	0.847	0.805	0.829	0.834	0.816
	CAS	0.576	0.557	0.628	0.588	0.611	0.607	0.594
λVAC-Ecore	Precision	0.804	0.792	0.847	0.817	0.842	0.841	0.819
	CAS	0.442	0.429	0.514	0.451	0.507	0.483	0.465
LOCLU	Precision	0.866	0.849	0.911	0.879	0.901	0.901	0.867
	CAS	0.601	0.584	0.657	0.611	0.638	0.642	0.623
IPCSO	Precision	**0.899**	**0.879**	**0.931**	**0.909**	**0.916**	**0.929**	**0.917**
	CAS	**0.647**	**0.632**	**0.694**	**0.651**	**0.617**	**0.689**	**0.671**

定位与查询节点相似的社区成员. TSCM性能表现虽好,但不足以与IPCSO方法竞争,这是由于IPCSO对社区内聚性的要求更高. 由于结构内聚性的要求,ACQ和VAC-Ecore在所有数据集上的表现均优于目标社区发现方法,且ACQ的CAS高于VAC-Ecore但却低于IPCSO,原因在于ACQ需要用户将查询节点所携带的属性作为输入并返回覆盖该属性的社区,而VAC-Ecore只需要找到属性得分最小的查询节点所在社区. 对于IPCSO,其返回的社区均在属性子空间的指导下,社区内节点与查询节点具有较大的相似性. LOCLU可与IPCSO方法相媲美,其在所有数据集上的表现均优于目标社区发现方法和基于结构度量的方法,但该方法容易将其他社区节点囊括在内. 整体而言,IPCSO始终优于所有方法.

4.5 案例分析(问题3)

本节将对比IPCSO和FocusCO在Disney上的运行结果,深入挖掘实验现象的原因. 图6给出了Disney数据集上两种方法的运行结果:

以影片ID是52和24的节点为查询节点,使用IPCSO与FocusCO方法分别搜索与v_{52}和v_{24}具有相似评分和价格的其他节点. 结果如图6所示. 其中红色节点v_{52}和v_{24}表示用户给定的查询节点,绿色节点表示离群点,算法找到与查询节点相似的节点用蓝色标识. 从图中可观察到,FocusCO检测到三个目标社区,而IPCSO找到了四个社区. IPCSO发现的每个社区都与用户偏好密切相关,并且精准的定位到了离群点,而FocusCO由于较少的查询节点使得算法错误的判别了离群点. 此外,FocusCO发现的社区的内部连接比IPCSO所发现的社区连接松散. 这是因为较少的查询节点对于FocusCO方法而言具有较大的局限性,而本文方法对于查询节点的个数不敏感,可通过较少的查询节点准确的捕获到用户偏好,从而使得社区搜索结果更精确.



(a) FocusCO (b) IPCSO

图6 IPCSO与FocusCO在Dinsey上的结果对比

5 结论

在属性网络中结合用户偏好的社区搜索和离群点检测任务中,基于现有大多数方法仅利用拓扑结构信息且只定位查询节点所在社区,未整合用户偏好到社区搜索过程中的局限性,设计了面向属性网络的融合用户偏好的多社区和离群点检测方法. 具体地,通过编码节点属性和结构得到候选查询节点集,利用候选查询节点集所携带的属性为核心属性设计了平均划分相似度来挖掘符合用户偏好的属性子空间方法,并设置结构和属性约束以搜索内聚社区. 真实数据集和人工数据集上的实验证明了本文方法的有效性.

参考文献

[1] JIN D, YOU X X, LI W H, et al. Incorporating network embedding into Markov random field for better community detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 160-167.

[2] 陈洁, 李锐, 赵姝, 等. 面向图表示社区检测的新型聚类覆盖算法[J]. 电子学报, 2020, 48(9): 1680-1687.
CHEN J, LI R, ZHAO S, et al. A new clustering cover algorithm based on graph representation for community detection[J]. Acta Electronica Sinica, 2020, 48(9): 1680-1687. (in Chinese)

[3] MATSUGU S, SHIOKAWA H, KITAGAWA H. Fast algorithm for attributed community search[J]. Journal of Information Processing, 2021, 29: 188-196.

[4] 潘剑飞, 董一鸿, 陈华辉, 等. 基于结构紧密性的重叠社区发现算法[J]. 电子学报, 2019, 47(1): 145-152.
PAN J F, DONG Y H, CHEN H H, et al. The overlapping community discovery algorithm based on compact structure[J]. Acta Electronica Sinica, 2019, 47(1): 145-152. (in Chinese)

[5] 马慧芳, 陈海波, 赵卫中, 等. 融合标签平均划分距离和结构关系的微博用户可重叠社区发现[J]. 电子学报, 2018, 46(11): 2612-2618.
MA H F, CHEN H B, ZHAO W Z, et al. Leveraging tag mean partition distance and social structure for overlapping microblog user community detection[J]. Acta Electronica Sinica, 2018, 46(11): 2612-2618. (in Chinese)

[6] LEE J Y, LEE J. Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction[J]. PLoS One, 2013, 8(4): e60372.

[7] FANG Y X, HUANG X, QIN L, et al. A survey of community search over big graphs[J].The VLDB Journal, 2020, 29 (1): 353-392.

[8] AGHAALIZADEH S, AFSHORD S T, BOUYER A, et al. A three-stage algorithm for local community detection based on the high node importance ranking in social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2021, 563: 125420.

[9] 於志勇, 陈基杰, 郭昆, 等. 基于影响力与种子扩展的重叠社区发现[J]. 电子学报, 2019, 47(1): 153-160.
YU Z Y, CHEN J J, GUO K, et al. Overlapping community detection based on influence and seeds extension[J]. Acta Electronica Sinica, 2019, 47(1): 153-160. (in Chinese)

[10] AKOGLU L, TONG H H, MEEDER B, et al. PICS: parameter-free identification of cohesive subgroups in large attributed graphs[C]//Proceedings of the 2012 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2012: 439-450.

[11] SHANG J W, WANG C K, WANG C P, et al. An attribute-based community search method with graph refining [J]. The Journal of Supercomputing, 2020, 76(10): 7777-7804.

[12] LIU H J, MA H F, CHANG Y, et al. Leveraging User Preferences for Community Search via Attribute Subspace [M]//Knowledge Science, Engineering and Management. Cham: Springer International Publishing, 2019: 584-595.

[13] FANG Y X, CHENG R, LUO S Q, et al. Effective community search for large attributed graphs[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1233-1244.

[14] LIU Q, ZHU Y F, ZHAO M J, et al. VAC: vertex-centric attributed community search[C]//2020 IEEE 36th International Conference on Data Engineering. Piscataway: IEEE, 2020: 937-948.

[15] YE W, MAUTZ D, BÖHM C, et al. Incorporating user's preference into attributed graph clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33 (12): 3716-3728.

[16] PEROZZI B, AKOGLU L, SÁNCHEZ P I, et al. Focused clustering and outlier detection in large attributed graphs [C]//KDD'14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA: ACM, 2014: 1346-1355.

[17] WU P, PAN L. Mining target attribute subspace and set of target communities in large attributed networks[EB/OL]. (2017-05-10). https://arxiv.org/abs/1705.03590.

[18] LI Y, SHA C F, HUANG X, et al. Community detection in attributed graphs: An embedding approach[J]. Proceedings of the AAAI Conference on Artificial Intelligence,

2018, 32(1): 338-345
[19] DE MÁNTARAS R L. A distance-based attribute selection measure for decision tree induction[J]. Machine Learning, 1991, 6(1): 81-92.
[20] 马慧芳, 张迪, 赵卫中, 等. 基于超图随机游走标签扩充的微博推荐方法[J]. 软件学报, 2019, 30(11): 3397-3412. MA H F, ZHANG D, ZHAO W Z, et al. Microblog recommendation method based on hypergraph random walk tag extension[J]. Journal of Software, 2019, 30(11): 3397-3412. (in Chinese)
[21] SEIDMAN S B. Network structure and minimum degree [J]. Social Networks, 1983, 5(3): 269-287.
[22] EBADIAN S, HUANG X. Fast algorithm for K-truss discovery on public-private graphs[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 2258-2264.
[23] WU J, LI C M, JIANG L, et al. Local search for diversified Top-k clique search problem[J]. Computers & Operations Research, 2020, 116: 104867.
[24] GIATSIDIS C, BERBERICH K, THILIKOS D M, et al. Visual exploration of collaboration networks based on graph degeneracy[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2012: 1512-1515.
[25] AYDIN K, BATENI M, MIRROKNI V. Distributed balanced partitioning *via* linear embedding[C]//WSDM'16: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2016: 387-396.

作者简介

李青青 女,1995年5月出生于甘肃省庆阳市. 现就读于西北师范大学计算机科学与工程学院. 主要研究方向为社区搜索.
E-mail: 2019211784@nwnu.edu.cn

马慧芳(通讯作者) 女, 1984年7月出生于甘肃省兰州市. 2010年毕业于中国科学院计算技术研究所. 现为西北师范大学计算机科学与工程学院教授, 主要研究方向为人工智能, 数据挖掘与机器学习.
E-mail: mahuifang@yeah.net

李 举 男, 1995年1月出生于山西省大同市. 现就读于西北师范大学计算机科学与工程学院. 主要研究方向为社区搜索.
E-mail: 2019221843@nwnu.edu.cn

李志欣 1971年10月出生, 广西桂林人. 2010年于中国科学院计算技术研究所博士研究生毕业, 获工学博士学位. 现为广西师范大学计算机科学与工程学院教授、博士生导师, 主要研究方向为图像理解、机器学习、跨媒体计算.
E-mail: lizx@gxnu.edu.cn

姜彦斌 男. 1996年1月出生于甘肃省陇南市. 现就读于西北师范大学计算机科学与工程学院. 主要研究方向为推荐系统, 图神经网络.
E-mail: jiangyanbin@nwnu.edu.cn