软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(3):1277−1291 [doi: 10.13328/j.cnki.jos.006792]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

属性公平的异质信息网络上的社区搜索算法*

乔连鹏[1], 侯会文[2], 王国仁[2]

[1](东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

[2](北京理工大学 计算机学院, 北京 100081)

通信作者: 乔连鹏, E-mail: qiaolp@stumail.neu.edu.cn

摘 要: 近年来, 异质信息网络上的社区搜索问题已经吸引了越来越多的关注, 而且被广泛应用在图数据分析工作中. 但是现有异质信息网络上的社区搜索问题都没有考虑子图上属性的公平性. 将属性的公平性与异质信息网络上的 kPcore 挖掘问题相结合, 提出了基于属性公平的异质信息网络上的极大 core 挖掘问题. 针对该问题, 首先提出了一个子图模型 FkPcore. 当对 FkPcore 进行枚举时, 基础算法 Basic-FkPcore 遍历了所有路径实例, 并枚举了大量 kPcore 及其子图. 为了提高算法效率, 提出了 Adv-FkPcore 算法, 以避免在枚举 FkPcore 时对所有的 kPcore 及其子图进行判断. 另外, 为了提高点的 *P_neighbor* 的获取效率, 提出了结合点标记的遍历方法(traversal method with vertex sign, TMS), 并基于 TMS 算法提出了 FkPcore 枚举算法 Opt-FkPcore. 在异质信息网络数据集上进行的大量实验证明了所提方法的有效性和效率.

关键词: 社区搜索; 异质信息网络; 属性公平性; 遍历方法; 枚举算法

中图法分类号: TP311

中文引用格式: 乔连鹏, 侯会文, 王国仁. 属性公平的异质信息网络上的社区搜索算法. 软件学报, 2023, 34(3): 1277−1291. http://www.jos.org.cn/1000-9825/6792.htm

英文引用格式: Qiao LP, Hou HW, Wang GR. Community Search Algorithm on Heterogeneous Information Networks Based on Attribute Fairness. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1277−1291 (in Chinese). http://www.jos.org.cn/1000-9825/6792.htm

Community Search Algorithm on Heterogeneous Information Networks Based on Attribute Fairness

QIAO Lian-Peng[1], HOU Hui-Wen[2], WANG Guo-Ren[2]

[1](College of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

[2](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: In recent years, community search on heterogeneous information networks has attracted more and more attention and has been widely used in graph data analysis. Nevertheless, the existing community search problems on heterogeneous information networks do not consider the fairness of attributes on subgraphs. This work combines attribute fairness with kPcore mining on heterogeneous information networks and proposes a maximum core mining problem on heterogeneous information networks based on attribute fairness. To solve this problem, a subgraph model called FkPcore is proposed. When enumerating FkPcore, the basic algorithm called Basic-FkPcore traverses all path instances and enumerates a large number of kPcores and their subgraphs. In order to improve the efficiency of the algorithm, an Adv-FkPcore algorithm is proposed to avoid judging all kPcores and their subgraphs when enumerating FkPcores. In addition, in order to improve the acquisition efficiency of P_neighbor, a traversal method with vertex sign (TMS) and a FkPcore enumeration algorithm called Opt-FkPcore based on the TMS algorithm are proposed. A large number of experiments on heterogeneous information networks demonstrate the effectiveness and efficiency of the proposed method.

* 基金项目: 国家自然科学基金(61732003, 61729201)

本文由“大数据治理的理论与技术”专题特约编辑杜小勇教授、杨晓春教授和童咏昕教授推荐.

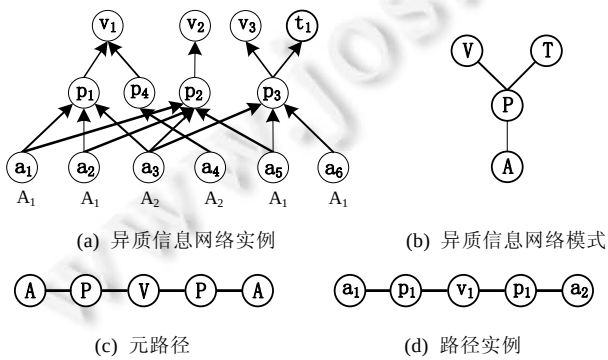
收稿时间: 2022-05-15; 修改时间: 2022-07-29; 采用时间: 2022-09-23; jos 在线出版时间: 2022-10-27

Key words: community search; heterogeneous information network; fairness of attributes; traversal method; enumeration algorithm

现实生活中, 很多的网络都是由组件及组件间的相互关系构成的, 比如社交网络、生物信息网络、金融关系网络及计算机网络等, 而这些网络都可以被统称为信息网络[1]. 近些年, 信息网络已经引起了多个领域专家的高度关注, 诸如生物医药领域、金融领域及计算机领域等. 这些领域的专家都积极地从信息网络中挖掘重要信息, 以开展蛋白质相互作用关系挖掘、金融风险防控和社交网络分析等方面的工作. 但是, 当前的挖掘工作主要都是基于同质信息网络(homogeneous information network)[2,3]的. 即网络中的对象和链接关系都是相同类型的. 比如, 朋友关系网络中仅仅包含人物个体和朋友关系, 这个网络中的对象和链接关系都是相同类型的. 同质信息网络通过忽略对象和关系的某些特性以达到简化网络的目的, 但是生活中的很多网络中都包含了多种类型的对象和关系, 就比如文献数据库, 其中包含了作者、论文、会议或期刊、时间等对象, 以及写、发表在(会议或期刊)、举办在或发表在(时间)等关系, 对这类信息网络, 就无法简单地基于同质信息网络进行高质量的网络分析.

近年来, 更多的研究者选择将这种信息网络构建为异质信息网络(heterogeneous information network, HIN), 并基于异质信息网络开展了诸多研究[4,5]. 异质信息网络是指对象及链接关系类型不止一种的特殊信息网络, 基于这个特性, 相较于同质信息网络, 异质信息网络可以包含更多的语义和结构信息, 这为数据挖掘领域提供了更高效的信息建模工具, 领域内专家利用这一工具实现了对现实网络更深层次更全面的信息挖掘工作. 就比如社交网络推荐问题, 就不再是简单地结合好友关系进行推荐, 也可以通过相似的兴趣、工作性质、故乡、浏览记录等进行更大范围的潜在好友挖掘与推荐. 这不仅能扩大推荐范围, 更能通过更多类型的对象与关系实现更精细的好友推荐. 相比之前基于同质信息网络的推荐工作, 基于异质信息网络的工作无疑具有更高的准确率[6−8].

异质信息网络其实就是具有多种对象和关系类型的网络, 当下, 这种模型已经被广泛用于生物信息网络、社交媒体网络和知识网络等的建模. 如图 1(a)所示, 这就是结合文献数据库信息建模成的一个异质信息网络, 这个网络包含了不同类型实体间的关系. 在这个网络中, 点标签为$\{a_1,a_2,a_3,a_4,a_5,a_6\}$, $\{p_1,p_2,p_3,p_4\}$, $\{v_1,v_2,v_3\}$, $\{t_1\}$的点分别表示作者、论文、会议、时间. 这个网络中包含了 4 种类型的实体对象, 而图中的有向边则表示实体间的关系. 比如, 点 a_1, a_2, p_1, v_1和它们之间的边表示作者 a_1和 a_2共同在会议 v_1上发表了论文 p_1. 而为了便于更好地理解异质信息网络, 人们提出了网络模式这一概念来对异质信息网络的元结构进行描述. 图 1(b)就是图 1(a)的异质信息网络模式, 图中的 *A*, *P*, *V*, *T* 分别表示点的对象类型.



(a) 异质信息网络实例 (b) 异质信息网络模式

(c) 元路径 (d) 路径实例

图 1 文献数据的异质信息网络

当前, 网络社区搜索[6−24]的相关工作其实可以被更细致地分为两类, 分别是社区发现[9−14]和社区搜索[15−19]: 社区发现问题旨在挖掘图中的所有社区, 而社区搜索问题则是结合给定的查询点来寻找包含该点的最具有凝聚力的子图. 本文研究的是异质信息网络上的社区搜索问题[17−19], 这个问题是网络信息挖掘的基本问题, 并且吸引了越来越多的关注. 但是已有的工作都没有考虑到异质信息网络中包含的一个重要信息, 即

社区中节点的属性公平性. 近些年来, 考虑到公平概念[25−37]的研究领域是机器学习领域, 人们发现, 机器学习训练在涉及性别、种族等与人类个体相关的敏感属性时, 往往会由于同机型偏差、算法本身抑或是个人偏见而引入歧视性行为. 因此, 为了消除这种歧视问题, 进一步改进机器学习效果而提出了一系列方法, 包括改进算法降低对敏感属性的依赖、制定相应指标来量化歧视程度等. 而在异质信息网络的社区搜索问题中, 我们认为也有必要引入公平性这一概念. 因为社会上由于性别、户籍差别带来的偏见屡见不鲜, 如果我们在社区搜索时, 能结合这些敏感属性找到一个或几个属性值完全相同的社区, 那么势必能为克服社区搜索结果的性别或户籍偏见带来极大的突破.

在本文工作中, 我们主要研究异质信息网络上的基于属性公平的社区搜索问题. 给定一个异质信息网络 G、一个查询点 $q\in G$ 和一个属性集 $AT=\{A_1,A_2,\ldots,A_m\}$, 我们的目标就是从 G 中找到一个包含点 q 的点集, 而点集中的其他点都具有和 q 一样的点类型, 并且所有点就给定的属性集 AT 满足属性公平的要求. 另外, 社区中点与点之间的连接是基于元路径(meta-path)的概念来定义的. 元路径已经在文献[38]中被研究过了, 它其实就是指一个连接两个给定点类型的点类型与边类型的序列(如图 1(c)所示). 在本文中, 我们首先结合文献[39]提出的 basic(k,P)-core 模型和公平性概念定义了一个新模型——FairkPcore(FkPcore)来表征异质信息网络上的公平社区. 一个异质信息网络上的公平社区 FkPcore 满足以下几点: (1) FkPcore 是一个极大子图; (2) FkPcore 包含查询点 q, 而且其他节点具有与 q 相同的节点类型; (3) FkPcore 中所有点属性 AT 的属性值集合完全涵盖给定的属性集 $AT=\{A_1,A_2,\ldots,A_m\}$的所有属性值, 并且每个属性值对应的节点数相等. 为了解决 FkPcore 的枚举问题, 我们提出了 3 种算法, 分别为 Basic-FkPcore, Adv-FkPcore 和 Opt-FkPcore. 其中, 基础算法 Basic-FkPcore 在枚举 FkPcore 时不仅遍历了图中所有的 path 实例, 而且枚举了大量非极大 kPcore. 为了提高算法的效率, 提出了 Adv-FkPcore 算法, 避免在枚举 FkPcore 时对所有的 kPcore 进行判断. 此外, 为了提高点的 *P_neighbor* 的获取效率, 结合点标记提出了新型遍历方法(traversal method with vertex sign, TMS), 并基于该方法提出了新的 FkPcore 枚举算法 Opt-FkPcore. 本文的主要贡献总结如下:

(1) 针对异质信息网络社区搜索问题, 提出了一个基于公平性概念的语义模型 FkPcore.
(2) 提出了一个基础算法 Basic-FkPcore, 来对异质信息网络中的 FkPcore 进行枚举.
(3) 结合剪枝技术, 设计了改进算法 Adv-FkPcore, 来对无用的遍历过程进行提前剪枝.
(4) 针对点的 *P_neighbor* 的搜索过程, 提出了基于点标记的遍历方法 TMS, 并优化了 FkPcore 枚举算法, 设计了算法 Opt-FkPcore.
(5) 在 4 个真实异质信息网络数据集上进行了一系列实验, 并证明了我们所提方法的有效性和效率.

本文第 1 节介绍相关工作. 第 2 节介绍与问题相关的术语及定义. 第 3 节介绍 FkPcore 枚举算法. 第 4 节对实验结果进行分析. 第 5 节对全文进行总结.

1 相关工作

- 社区搜索

社区搜索问题旨在寻找所有包含给定查询点的稠密子图. 而在对这些子图的凝聚性进行定义时, 大多数工作采用的是最小度数这一度量手段, 它要求子图中的每个点的度数都大于等于 k, 这与 k-core 的概念是一致的[40−42]. 比如, Sozio 等人[43]结合 k-core 提出了包含查询点的语义模型来解决社区搜索问题, Zhang 等人[44]则结合 k-core 解决了属性图上的以关键词为中心的社区搜索问题, 而其他被用在社区搜索问题中的语义模型还包括 k-truss[45−47], k-clique[16,48]以及 K-ECC 等. 比如, Huang 和 Chen 等人结合 k-truss 来解决社区搜索问题[15,49], Yuan 等人[48]提出了一种基于 k-clique 的 k-clique 渗透社区模型来解决最密集 clique 渗透社区的搜索问题. 以上这些工作都是基于同质信息网络的, 而 Fang 等人[39]则是做了异质信息网络上的第一个社区搜索问题. Fang 等人结合 k-core 提出了 3 种异质信息网络上的新模型, 分别是 basic(k,P)-core, edge-disjoint(k,P)-core 以及 vertex-disjoint(k,P)-core, 并设计了相应的查询算法及索引技术来解决异质信息网络上的社区搜索问题. 但是不管是同质信息网络还是异质信息网络的社区搜索问题, 它们都没有考虑到设计敏感属性的公平性这一点.

- 基于公平的数据挖掘

属性, 特别是敏感属性的公平性不仅是机器学习领域, 也是图数据挖掘领域一个十分重要的概念. 但是, 当前大部分考虑属性公平性的工作都是在机器学习领域的. Zehlike 等人[50]提出了一种通过生成排名来确保团体公平性的方式, 可以保证排名中敏感属性的所占比例不低于某个给定阈值. Serbos 等人[51]针对组推荐中的公平性问题, 提出了一种贪心算法来寻找近似解. Beutel 等人[52]则在研究推荐系统的公平性问题时, 提出了一系列的指标来评估算法的公平性. 而在图数据挖掘领域, 第一个涉及属性公平性的稠密子图挖掘工作是 Pan 等人[53]结合属性公平性针对极大团枚举问题开展的研究. 由此可见, 越来越多的领域开始将目光投向了属性公平性这一重要概念. 但是, 当下并没有在异质信息网络上开展任何结合属性公平性的社区搜索问题的研究, 异质信息网络由于可以融合异构信息源的信息, 进而可以全面刻画用户特征, 但却依然无法从敏感属性上克服由于统计的属性值偏差带来的偏见. 因此, 在异质信息网络上开展结合属性公平性的社区搜索问题势在必行.

2 基本概念

在这一节, 我们介绍了本文涉及的几个重要定义, 并进一步对基于属性公平的异质信息网络上的社区搜索问题做了定义. 表 1 罗列了本文在问题定义和内容阐述过程中用到的主要符号.

表 1 符号表

符号	符号说明
φ	对象类型映射函数
ψ	关系类型映射函数
$\mathcal{A}$	对象类型集合
$\mathcal{R}$	关系类型集合
G	$G(V,E,\varphi,\psi)$, 一个异质信息网络
$T_G(\mathcal{A},\mathcal{R})$	异质信息网络 G 的网络模式
P	元路径
$\deg(v,S)$	点 v 在点集 S 中的 *P_neighbor* 数
AT	给定的属性值集合
$\deg_{\max}(v,S)$	S 中所有 $\deg(v,S)$的最大值

定义 1(异质信息网络[4,5]). 异质信息网络其实就一个有对象类型映射函数$\varphi:V\to\mathcal{A}$和关系类型映射函数$\psi:E\to\mathcal{R}$的有向图 $G(V,E,\varphi,\psi)$. 其中, 每一个点 $v\in V$ 都满足$\varphi(v)\in\mathcal{A}$, 每一条边 $e\in E$ 都满足$\psi(e)\in\mathcal{R}$, 且满足$|\mathcal{A}|>1$ 或者$|\mathcal{R}|>1$.

定义 2(网络模式[1,11]). 给定一个异质信息网络 $G(V,E,\varphi,\psi)$, G 的网络模式就是一个以$\mathcal{A}$中的点类型为点、以$\mathcal{R}$中的边类型为边的有向图 $T_G(\mathcal{A},\mathcal{R})$. 另外, 考虑到本文所研究的异质信息网络中的关系均为对称关系, 因此, 本文的异质信息网络的网络模式是用无向图来表示的.

异质信息网络的网络模式对多种点类型与边类型的连接情况进行了概括, 网络模式中的一条边可能代表了一对一、一对多甚至是多对多的关系. 图 1(b)就是图 1(a)的异质信息网络模式, 图中的 A, P, V, T 分别表示点的对象类型. 另外, 对于目标类型 A 到目标类型 P 的关系类型 R, 应表示为 $A\xrightarrow{R}P$, 其中, A 和 P 分别为关系 R 的源对象类型以及目标对象类型, 那么自然就有逆关系 R^{-1}. 本文所研究的异质信息网络中的关系均为对称关系, 因此 R 等于 R^{-1}.

定义 3(元路径[5]). 元路径 P 是在网络模式 $T_G(\mathcal{A},\mathcal{R})$上定义的路径, 可以表示为 $A_1,A_2,\ldots,A_{l+1}$, 并且满足$A_i\in\mathcal{A}(1\leqslant i\leqslant l+1)$以及$\psi(A_j,A_{j+1})\in\mathcal{R}(1\leqslant j\leqslant l)$.

图 1(c)就是异质信息网络图 1(a)的一个元路径, 对于元路径, 我们也可以用点类型序列来表示, 就比如图 1(c)的元路径就可以表示为 $P=(APVPA)$.

定义 4(路径实例). 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$, 路径 $a_1a_2\ldots a_{l+1}$ 如果满足$\varphi(a_i)=A_i(1\leqslant i\leqslant l+1)$以及$\psi(a_j,a_{j+1})=\psi(A_j,A_{j+1})(1\leqslant j\leqslant l)$, 则该路径就被称为元路径 P 的一个路径实例.

定义 5(*P_neighbor*). 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$, 对于 G 中任意两点 u 和 v 来说, 如果 u 和 v 之间存在一条路径, 并且该路径是元路径 P 的一个路径实例, 那么 u 和 v 互为 *P_neighbor*.

例如, 图 1(d)就是图 1(c)的元路径的一个路径实例, 而图中的点 a_1 和 a_2 分别是彼此的 *P_neighbor*. 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$、一个查询节点 q, 对于一个包含所有与 q 点类型一致的节点的集合 S 来说, 集合中每个点 v 的度数 $\deg(v,S)$等于 v 在 S 中的 *P_neighbor* 数.

定义 6(kPcore[39]). 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$、一个整数 k、一个查询节点 q、大小阈值 *min_size*, kPcore 就是一个点数大于等于 *min_size* 的极大子图 S, 子图中的其他点的点类型与 q 一致, 并且 S 中的每个点均满足 $\deg(v,S)=4\geqslant k$.

以图 1(a)为例, 取 $P=(APVPA)$, k 为 3, q 为 a_1. 取 $S=\{a_1,a_2,a_3,a_4,a_5,a_6\}$, 则 $\deg(a_1,S)=4$, $\deg(a_2,S)=4$, $\deg(a_3,S)=5$, $\deg(a_4,S)=4$, $\deg(a_5,S)=4$, $\deg(a_6,S)=2$, 点 a_6 不满足度数要求, 将其从 S 中移除, 剩余点的度数分别为 $\deg(a_1,S)=4$, $\deg(a_2,S)=4$, $\deg(a_3,S)=4$, $\deg(a_4,S)=4$, $\deg(a_5,S)=3$, 则点集$\{a_1,a_2,a_3,a_4,a_5\}$就是我们要找的一个 kPcore.

定义 7(FkPcore). 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$、一个整数 k、一个查询节点 q、大小阈值 *min_size*、查询点 q 上某个属性值对应属性的所有值的集合 $AT=\{A_1,A_2,\ldots,A_m\}$, FkPcore 就是 G 上一个包含查询节点 q 的 kPcore, 并且满足: (1) 所有点的属性值集合涵盖 AT 中的所有值; (2) 所有属性值对应的点数相等.

还是以图 1(a)为例, 取 $P=(APVPA)$, $AT=\{A_1,A_2\}$, k 为 3, q 为 a_1, 则点集$\{a_1,a_2,a_3,a_4\}$就是我们要找的其中一个 FkPcore.

问题定义. 给定一个异质信息网络 $G(V,E,\varphi,\psi)$、一个元路径 $P=(A_1,A_2,\ldots,A_{l+1})$、一个整数 k、一个查询节点 q、大小阈值 *min_size*、查询点 q 上某个属性值对应属性的所有值的集合 $AT=\{A_1,A_2,\ldots,A_m\}$, 异质信息网络上的公平社区搜索问题就是找到 G 中所有的 FkPcore.

引理 1. 给定一个异质信息网络 G、查询点 q、元路径 P、属性集 $AT=\{A_1,A_2,\ldots,A_m\}$、core 值 k, 对包含 q 的 kPcoreS 和任意 FkPcoreC, 两者肯定满足 $C\subseteq S$.

证明: 假设存在一个包含 q 的 kPcoreS 和一个包含 q 的 FkPcoreC, 并且不满足 $C\subseteq S$, 那么至少存一个点 $v\in C$ 不被 S 包含. 但是 $v\in C$ 表明, v 在 C 中的度数大于等于 k. 这又与 $v\notin S$ 相违背, 因此假设不成立. 得证. □

3 基本求解过程

算法 1 是我们提出的基础算法, 在该算法中, 我们对所有包含查询点 q 的 kPcore 进行枚举, 然后结合属性集 AT 对 FkPcore 进行挖掘. 算法的具体过程如下所示.

算法 1. Basic-FkPcore.

输入: 异质信息网络图 G、元路径 P、属性集 $AT=\{A_1,A_2,\ldots,A_m\}$、查询点 q、core 值 k、大小阈值 *min_size*.

输出: FkPcore 集 R.

1. $S\leftarrow\{v$ 的点类型与 q 一致$\}$;
2. $\deg(v,S)\leftarrow\{v$ 在 S 中的 *P_neighbor* 数$\}$;
3. **for** S 中的每个点 v **do**
4. **if** $\deg(v,S)<k$ **then** $Q.push(v)$;
5. **while** $(Q\neq\varnothing)$ **do**
6. $v\leftarrow Q.pop(\cdot)$;
7. remove v from S;
8. **for** S 中的每个点 u **do**
9. **if** $u\in P\_neighbor[v]$ **then**

```
10.          deg(u,S)−−;
11.          if deg(v,S)<k then Q.push(u);
12. kPcoreEnum(R,deg,P_neighbor,S,min_size);
13. for R 中的每个点集 C do
14.     if C 不包含查询点 q OR C 不满足公平条件 Or R 中存在点集能包含 C then remove C from R;
15. return R;
16. Procedure kPcoreEnum(R,deg,P_neighbor,S,min_size)
17. if S=∅ OR S<min_size then return;
18. S′=S;
19. for S′中的每个点 v do
20.     remove v from S′;
21.     for P_neighbor[v]中的每个点 u do
22.         if deg(v,S′)>k then deg(u,S′)−−;
23.         else return;
24.     if |S′|≥min_size then R.push(S′);
25.     kPcoreEnum(R,deg,P_neighbor,S′,min_size);
```

算法 1 的第 1、2 行是通过遍历全图获取图 G 中与目标点类型一致的点的集合 S 以及 S 中所有点在 S 中的 *P_neigbor* 数(即度数), 这一步涉及对全图 path 实例的完全遍历, 计算代价大.

算法 1 的第 3–11 行是对不满足度数约束的点进行剪枝, 缩小遍历规模, 提高算法效率.

算法 1 的第 12–15 行是对 G 中的 FkPcore 进行挖掘, 最终返回包含点 q 的 FkPcore 集 R.

算法 1 的第 16–25 行是对剪枝后的点集 S 中的所有满足度数约束和 *min_size* 约束的子图进行枚举,以方便找到所有包含点 q 的 FkPcore.

算法 1 中的 kPcoreEnum 过程挖掘出了所有由与查询点 q 点类型一致的节点组成的子图, 该子图满足度数约束和 *min_size* 约束, 然后对每个子图进行 FkPcore 的判定. 这样的计算代价是很大的.

结合算法内容, 可以得到算法 1 的时间复杂度为 $O\left(m+\deg_{\max}(v,S)\cdot\sum C_{|S|}^{i}\right)$, 其中, $\deg_{\max}(v,S)$表示 S 中所有点度数的最大值, $|min\_size|\leqslant i\leqslant|S|$.

以图 1(a)为例, 取 P=(*APVPA*), AT={A_1,A_2}, k 为 3, q 为 a_1, *min_size*=4. 取与查询点类型一致的点集合 S={a_1,a_2,a_3,a_4,a_5,a_6}, 则 deg(a_1,S)=4, deg(a_2,S)=4, deg(a_3,S)=5, deg(a_4,S)=4, deg(a_5,S)=4, deg(a_6,S)=2. 由于点 a_6 不满足度数要求, 因此将其从 S 中移除, S 中剩余点的度数分别为 deg(a_1,S)=4, deg(a_2,S)=4, deg(a_3,S)=4, deg(a_4,S)=4, deg(a_5,S)=3, 从当前点集 S 中找到所有子集分别为: {a_1,a_3,a_4,a_5}, {a_1,a_2,a_4,a_5}, {a_1,a_2,a_3,a_5}, {a_1,a_2,a_3,a_4},…, 然后结合算法第 13–15 行, 确定点集{a_1,a_2,a_3, a_4}就是一个 FkPcore.

针对算法 1 计算代价大这一点, 我们发现, 在 kPcoreEnum 过程中挖掘出来的所有子图, 其中很可能存在某些子图直接满足 FkPcore 的条件, 而那些被这些子图所包含的其他子图便没有必要再进行 FkPcore 的判断. 我们将几个发现整理成了如下几个定理, 以尽可能提前终止判定过程.

引理 2. 给定一个异质信息网络 G、查询点 q、元路径 P、属性集 AT={A_1,A_2,…,A_m}、core 值 k、大小阈值 *min_size*,对于任何一个包含点 q 的目标类型节点集合 S, 如果满足$|S|$<max(*min_size*,$|AT|_{\min}$), 其中, $|AT|_{\min}$ 表示$|AT|$的倍数中大于 k+1 的最小值, 那么 S 不可能是 FkPcore.

证明: 对于点集 S, 如果$|S|$<*min_size*, 则 S 肯定不是 FkPcore; 如果$|S|$<$|AT|_{\min}$, 则 S 不可能在所有点度数大于等于 k 的情况下满足属性的公平性要求, 也就意味着 S 肯定也不是 FkPcore. 得证. □

引理 3. 给定一个异质信息网络 G、查询点 q、元路径 P、属性集 AT={A_1,A_2,…,A_m}、core 值 k、大小阈值 *min_size*,对于任何一个包含点 q 的目标类型节点集合 S, 将 S 中的所有点按照其属性值 A_i 分配到对应的集合

F_{A_i} 中去.对于这些集合, 如果存在一个集合 F_{A_i} 满足 $|F_{A_i}|<\max\left(\frac{min\_size}{|AT|},\frac{|AT|_{\min}}{|AT|}\right)$, 那么 S 不可能是 FkPcore.

证明: 如果存在某个 $|F_i|<\frac{min\_size}{|AT|}$, 那么找到的任何点集 S 都不可能在满足属性公平性的前提下保证 $S\geqslant min\_size$; 另一方面, 如果存在 $|F_i|<\frac{|AT|_{\min}}{|AT|}$, 那么任何点集 S 都不可能在满足所有点度数大于等于 k 的前提下满足属性的公平性要求. 这两种情况下, S 均不可能是 FkPcore. 得证. □

引理 4. 给定一个异质信息网络 G、查询点 q、元路径 P、属性集 $AT=\{A_1,A_2,\ldots,A_m\}$、core 值 k、大小阈值 min_size,对于任意两个满足 core 值和 min_size 约束、并且包含点 q 的子图 S 和 X 来说, 假设存在 $X\subseteq S$, 如果 S 是一个 FkPcore, 那么 X 肯定不是 FkPcore.

证明: 如果 S 是一个 FkPcore, 那么 S 的任何子集都不可能是个 FkPcore, 因为 S 的所有子集都不满足极大性的要求. 得证. □

结合上述内容, 我们提出了改进后的 FkPcore 挖掘算法——Adv-FkPcore. 算法 2 的具体过程如下所示.

算法 2. Adv-FkPcore.

输入: 异质信息网络图 G、元路径 P、属性集 $AT=\{A_1,A_2,\ldots,A_m\}$、查询点 q、core 值 k、大小阈值 min_size.

输出: FkPcore 集 R.

1. $S\leftarrow\{v$ 的点类型与 q 一致$\}$;
2. $\deg(v,S)\leftarrow\{v$ 在 S 中的 $P\_neighbor$ 数$\}$;
3. **for** S 中每个点 v **do**
4. **if** $\deg(v,S)<k$ **then** $Q.push(v)$;
5. **while** $(Q\neq\varnothing)$ **do**
6. $v\leftarrow Q.pop(\cdot)$;
7. remove v from S;
8. **for** S 中的每个点 u **do**
9. **if** $u\in P\_neighbor[v]$ **then**
10. $\deg(u,S)--$;
11. **if** $\deg(v,S)<k$ **then** $Q.push(u)$;
12. **for** S 中每个点 v **do**
13. **if** v 与 p 之间不存在 path 实例相连 **then** remove v from S;
14. **if** S 满足属性公平性要求 **then return** S;
15. **if** $|S|<\max(min\_size,|AT|_{\min})$ **then return** $\varnothing$;
16. 将 S 中所有点 v 按照其属性值 A_i 分配到对应的集合 F_{A_i} 中去;
17. **for** all $i=0,1,\ldots,|AT|$ **do**
18. **if** $|F_i|<\max\left(\frac{min\_size}{|AT|},\frac{|AT|_{\min}}{|AT|}\right)$ **then return** $\varnothing$;
19. $FkPcoreEnum(R,\deg,P\_neighbor,S,min\_size)$;
20. **return** R;
21. Procedure $FkPcoreEnum(R,\deg,P\_neighbor,S,min\_size)$
22. **if** $S=\varnothing$ OR $|S|<min\_size$ OR $q\notin S$ **then return**;
23. **for** S 中的每个点 v **do**
24. $S'\leftarrow S-\{v\}$;
25. **for** $P_neighbor[v]$中的每个点 u **do**

26. **if** deg(u,S')>k **then** deg(u,S')−−;
27. **else** continue;
28. **if** $|S'|\geq min\_size$ AND S'满足公平条件 **then** R.*push*(S'); break;
29. *FkPcoreEnum*(R,deg,$P\_neighbor$,S',min_size);

算法 2 的第 3–13 行是利用引理 1 对图 G 进行剪枝, 这极大缩小了遍历范围.

算法 2 的第 14–18 行利用引理 2 和引理 3 对点集进行判断, 以提前终止对无用点集的判定过程.

算法 2 的第 21–29 行是 FkPcore 枚举过程, 该过程的第 28 行则是利用引理 4 对 FkPcore 的枚举过程进行剪枝, 以减少无用枚举过程, 提高枚举效率.

以图 1(a)为例, 取 P=($APVPA$), AT={A_1,A_2}, k 为 3, q 为 a_1, min_size=4. 取与查询点类型一致的点集合 S={a_1,a_2,a_3,a_4,a_5,a_6}, 则 deg(a_1,S)=4, deg(a_2,S)=4, deg(a_3,S)=5, deg(a_4,S)=4, deg(a_5,S)=4, deg(a_6,S)=2. 由于点 a_6 不满足度数要求, 因此将其从 S 中移除, 点集 S 中剩余点的度数分别为: deg(a_1,S)=4, deg(a_2,S)=4, deg(a_3,S)=4, deg(a_4,S)=4, deg(a_5,S)=3, 则点集{a_1,a_2,a_3,a_4,a_5}就是一个 kPcore. 但是该点集不满足属性公平性要求, 因此该点集不是一个 FkPcore. 移除 a_4, 则点集{a_1,a_2,a_3,a_5}不满足属性公平性要求. 移除 a_5, 得到 deg(a_1,S)=3, deg(a_2,S)=3, deg(a_3,S)=3, deg(a_4,S)=4, 则点集{a_1,a_2,a_3,a_4}就是一个 FkPcore.

结合算法内容, 可以得到算法 2 的时间复杂度为 $O(m+\deg_{\max}(v,S_i)\cdot\sum C_{|S_i|}^{|FkPcore_i|})$, 其中, $\deg_{\max}(v,S_i)$表示当前点集 S_i 中所有点度数的最大值, $FkPcore_i$ 则表示当前点集 S_i 的 FkPcore.

但是算法 2 在为所有具有目标类型的节点寻找 $P\_neighbor$ 时, 会将所有的 path 实例都枚举一遍, 这耗费了大量的计算时间. 另外, 通过观察发现, 并不是所有具有目标类型的节点都能构成 FkPcore. 因此, 如果为寻找每个点的 $P\_neighbor$ 而遍历所有 path 实例, 会造成大量资源浪费.

针对这个问题, 本文提出了结合点标记的遍历方法(traversal method with vertex sign, TMS), 用以获取点的 $P\_neighbor$ 集合. 该遍历方法的灵感是来自于元路径的对称特点, 由于元路径的对称性, 所以当去寻找某个节点 v 的 $P\_neighbor$ 时, 只需沿着元路径去判断一半的长度即可, 也就是说, 对于给定的元路径 P=($A_1A_2\ldots A_{l+1}$)和查询节点 q, 只需从节点 q 出发, 沿着元路径找到符合约束的节点 u, 这样就相当于确定了存在一条从查询节点出发的元路径 P 的实例. 具体过程如算法 3 所示.

算法 3. TMS.

输入: 异质信息网络图 G、元路径 P=($A_1\ldots A_{l+1}$)、查询点 q、core 值 k、$P\_neighbor$.

输出: 与查询点 q 点类型一致且彼此通过路径实例相连的点的 $P\_neighbor$.

1. S←与 q 相连且点类型为 A_2 的节点;
2. **for** S 中的每个节点 v **do**
3. **if** l=2 OR G 中存在一个点类型为元路径第 l/2+1 个点类型的节点 u 与 v 通过元路径 $A_2\ldots A_{l/2+1}$ 相连 **then**
4. $v.sign$←$visited$;
5. **for** S 中的每个节点 v **do**;
6. **if** $v.sign$=$visited$ **then**
7. $Cand$(v)←与节点 v 相连且点类型为查询节点类型的节点;
8. **for** $Cand$(v)中的每个点 u **do**
9. $P\_neighbor$[u]←$Cand$(v)中除节点 u 以外的所有点;
10. **return** $P\_neighbor$;

以图 2 为例, 取 P=($APVPA$), 查询节点为 a_1, 当从节点 a_1 出发沿着元路径途径节点 p_1 到达节点 v_1 时, 就可以确定一个从查询节点出发的路径实例, 此时则置节点 p_1 的 $sign$ 为 $visited$, 这表示已经找到了从查询节点出发的一个路径实例, 而通过该路径实例与查询节点相连的点, 则是与查询节点的第 1 个邻居节点也就是 p_1 直接相连的其他具有查询节点类型的节点, 也即节点 a_2 和 a_3, 这两个节点与查询节点互为彼此的

P_neighbor. 而后续再寻找查询节点的其他 *P_neighbor* 时, 就会跳过 *sign* 值为 *visited* 的节点. 结合这个例子可以发现通过这种方式可以极大地减少寻找 *P_neighbor* 时的遍历过程, 这有助于提高算法执行效率.

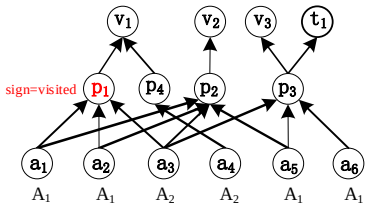


图 2 TMS 示例

结合上述 TMS 算法, 我们对算法 2 进行了优化, 进而得到了下面的 Opt-FkPcore 算法. 在算法 4 中, 我们利用 TMS 算法可以减少 path 实例的重复性遍历, 提高算法执行效率, 减少计算资源的浪费. 算法 4 的具体过程如下所示.

算法 4. Opt-FkPcore.

输入: 异质信息网络图 G、元路径 P、属性集 $AT=\{A_1,A_2,\ldots,A_m\}$、查询点 q、core 值 k、大小阈值 *min_size*.

输出: FkPcore 集 R.

1. *TMS*(G,P,q,k,*P_neighbor*);
2. $S\leftarrow$\{v 的点类型与 q 一致\};
3. deg(v,S)$\leftarrow$|*P_neighbor*[v]|;
4. **for** S 中每个点 v **do**
5. **if** deg(v,S)<k **then** *Q.push*(v);
6. **while** ($Q\neq\varnothing$) **do**
7. $v\leftarrow$*Q.pop*($\cdot$);
8. remove v from S;
9. **for** S 中的每个点 u **do**
10. **if** $u\in$*P_neighbor*[v] **then**
11. deg(u,S)$--$;
12. **if** deg(v,S)<k **then** *Q.push*(u);
13. **if** |S|<max(*min_size*,$|AT|_{\min}$) **then return** $\varnothing$;
14. **if** S 满足属性公平性要求 **then return** S;
15. 将 S 中所有点 v 按照其属性值 A_i 分配到对应的集合 F_{A_i} 中去;
16. **for** all i=0,1,…,|AT| **do**
17. **if** $|F_i|<\max\left(\frac{min\_size}{|AT|},\frac{|AT|_{\min}}{|AT|}\right)$ **then return** $\varnothing$;
18. *FkPcoreEnum*(R,deg,*P_neighbor*,S,*min_size*);
19. **return** R;
20. Procedure *FkPcoreEnum*(R,deg,*P_neighbor*,S,*min_size*)
21. **if** $S=\varnothing$ OR |S|<*min_size* OR $q\notin S$ **then return**;
22. **for** S 中的每个点 v **do**
23. $S'\leftarrow S-\{v\}$;
24. **for** *P_neighbor*[v]中的每个点 u **do**
25. **if** deg(u,S')>k **then** deg(u,S')$--$;
26. **else** continue;

27. **if** $|S'|\geq min\_size$ AND S'满足公平条件 **then** $R.push(S')$; break;

28. $FkPcoreEnum(R,\deg,P\_neighbor,S',min\_size)$;

算法 4 利用 TMS 算法提高了具有目标类型的节点的 *P_neighbor* 的挖掘效率, 减少了冗余计算, 这可以极大提高算法的执行效率.

4 实验分析

4.1 实验数据

我们在 4 个真实数据集上进行实验, 分别包括 Foursquare、DBLP、IMDB 和 DBpedia. Foursquare 有 5 种点类型, 包含了美国的用户签到信息. DBLP 有 4 种点类型, 包含了计算机科学领域的出版信息. IMDB 有 4 种点类型, 包含了自 2000 年以来的电影评分信息. DBpedia 有 4 种点类型, 包含了从 Wikipedia 抓取的信息. 表 2 给出了数据集所对应的详细信息.

表 2 实验数据集

数据集	点数	边数	点类型数	元路径数
Foursquare	43 199	405 476	5	20
DBLP	682 819	1 951 209	4	14
IMDB	4 467 806	7 597 591	4	6
DBpedia	5 900 558	17 961 887	413	50

4.2 实验环境配置

为了在异质信息网络中结合属性公平性进行社区的搜索, 我们结合本文设计的几种算法实现了 3 个对比方法, 分别是 BasicFkP、AdvFkP、OptFkP. BasicFkP 是结合算法 1 实现的 FkPcore 搜索方法. AdvFkP 是结合算法 2 实现的改进算法. OptFkP 则是结合 TMS 算法实现的优化方法. 所有的算法都是用 C++实现的. 实验的电脑配置为 Intel(R) Core(TM) i5-9500 CPU@3.00 GHZ 16 GB 内存, 电脑操作系统版本为 Windows 10 X64.

4.3 实验方法

在属性公平性方面, 我们选择以性别信息和国家信息作为实验中涉及的敏感信息, 而所有的实验也是基于这两个属性中的某一个进行的基于属性公平性的社区搜索. 在元路径方面, 我们收集了 4 个数据集的元路径信息, 如表 2 所示. 需要特别注意的是第 4 个数据集 DBpedia, 这个数据集由于规模比较大, 包含的点类型数较多, 所以这个数据集中包含的元路径信息十分丰富. 考虑现实生活中关联性越大的点连接彼此的元路径长度越短这一点, 我们选取了 DBpedia 中长度小于等于 4、频率排名前 50 的 50 条元路径[39]. 而另外 3 个数据集, 我们选择了 3 个数据集全部的元路径. 在进行查询时, 我们针对每一个数据集进行 40 次查询, 每次查询都是从元路径中随机选取一个, 然后按照度数从大到小选择一个点作为查询点(度数大于等于 6). 查询的默认属性如下所示, core 值 k=6, 敏感属性集为性别 AT={男,女}, min_size=7. 除非特殊声明, 否则后续实验的配置均为默认配置, 每个点对应的结果都是 40 次查询的平均值.

4.4 实验结果与分析

4.4.1 FkPcore 结果分析

为了对 FkPcore 模型进行分析, 我们首先针对 FkPcore 就不同 k 值的大小分布进行了实验. k 的取值范围为[0,100], 相关元路径分别为 P_1=*APVPA* 以及 P_2=*APTPA*. 具体实验结果如图 3 所示, 结合实验结果我们发现, 元路径 P_1 对应的 FkPcore 与 P_2 对应的 FkPcore 的结果规模比较接近. 这证明我们这个模型在不同实验参数下的结果凝聚性上表现出色.

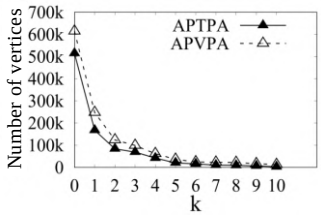


图 3 FkPcore 就不同 k 值的点数分布情况

4.4.2 案例分析

本节, 我们在数据集 DBLP 上进行了查询, 并结合结果进行案例分析. 本节一共进行了两次查询,查询参数分别为: (1) q=Daniel Genkin, P_1=*APVPA*, *AT*(性别)={男,女}, k=10, *min_size*=4; (2) q=Daniel Genkin, P_1=*APVPA*, *AT*(国家)={美国,奥地利}, k=10, *min_size*=4. 注意, 此处将 k 值设为 10, 是因为 k 值太小的话查询到的结果点数过多, 受空间限制, 我们决定将 k 值设为 10.

结果见表 3, 结合结果我们发现: 在其他参数一致的情况下, 敏感属性的差异会导致结果的明显变化. 放到实际生活中来看, 当我们不考虑敏感属性时, 得到的结果为 Daniel Genkin, Daniel Gruss, Mickael Schwarz, Mike Hambury, Moritz Lipp, Paul Kocher, Stefan Mangard, Thomas Prescher, Werner Haas, Yuval Yarom, Diego Gragnaniello, Francesco Marra, Giovanni Poggi, Limin Zhang, Lu Feng, Luisa Verdoliva, Pengyuan Lu. 从性别角度来看, 这个结果集合中仅仅存在两名女性专家, 放到现在这个强调男女平等的社会中来看, 这个结果明显存在性别上的偏见. 而从国家角度来看也一样, 该结果集国籍为美国的专家有 6 名, 而中国籍的只有 1 名, 从国籍角度看, 这也明显存在国籍上的歧视问题. 因此, 结合案例结果来看, 我们提出的 FkPcore 模型明显能为消除因性别、国籍等敏感信息差别而带来的歧视问题发挥辅助作用.

表 3 案例分析

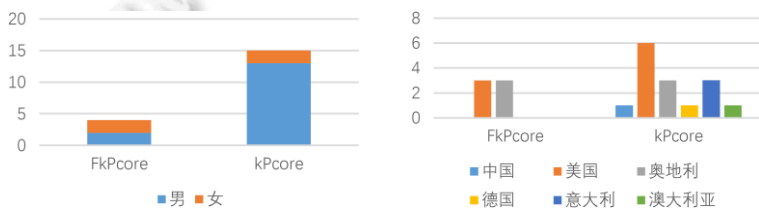
参数	结果
q=Daniel Genkin, P_1=*APVPA*, *AT*(性别)={男,女}, k=10, *min_size*=4	Daniel Genkin, Daniel Gruss, Lu Feng, Luisa Verdoliva
q=Daniel Genkin, P_1=*APVPA*, *AT*(国家)={美国,奥地利}, k=10, *min_size*=4	Daniel Genkin, Daniel Gruss, Mike Hambury, Moritz Lipp, Paul Kocher, Stefan Mangard

为了说明使用本文方法后, 结果在属性分布上的变化, 我们不考虑属性公平, 使用如下查询参数: q=Daniel Genkin, P_1=*APVPA*, k=10, *min_size*=4 来获取包含当前查询点的极大 kPcore, 结果见表 4.

表 4 案例分析(不考虑属性公平)

参数	结果
q=Daniel Genkin, P_1=*APVPA*, k=10, *min_size*=4	Daniel Genkin, Daniel Gruss, Mickael Schwarz, Mike Hambury, Moritz Lipp, Paul Kocher, Stefan Mangard, Yuval Yarom, Diego Grag- naniello, Francesco Marra, Giovanni Poggi, Limin Zhang, Lu Feng, Luisa Verdoliva, Pengyuan Lu

结合图 4 我们可以发现, 在考虑属性公平时, 我们得到的结果在属性分布上趋于平均. 结合现实情况, 我们提出的方法将有助于解决社区搜索结果中可能存在的性别或国籍偏见问题.



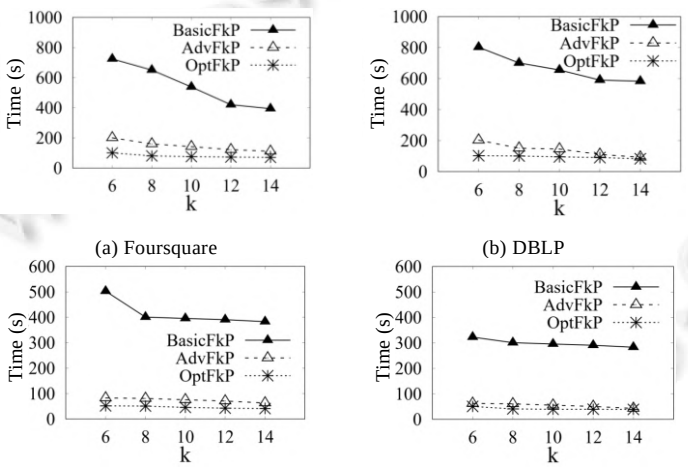
(a) *AT*(性别)={男,女} (b) *AT*(国家)={美国,奥地利}

图 4 属性分布变化

4.5 效率测试

4.5.1 运行时间测试

我们在表 2 所提到的 4 个数据集上对 BasicFkP、AdvFkP、OptFkP 这 3 种方法进行了运行时间的测试. 如图 5(a)所示, 尽管数据集 Foursquare 的点数和边数规模都不大, 但是这个数据集点的平均度数较大, 数据集稠密,所在这个数据集上运行那 3 种方法所消耗的时间都比较大. 而该数据集上, BasicFkP 方法在 k 大于 10 时运行时间出现了大幅度减少. 这是因为 k 值约束剪枝掉很多无用点, 缩减了遍历空间, 导致运行时间出现了明显降低. 而在数据集 Foursquare 上, 方法 AdvFkP 的运行时间远远低于 BasicFkP. 这是因为该方法结合了引理 1–引理 4, 不仅利用 FkPcore 与 kPcore 的关系(引理 1)来结合 kPcore 对无用点进行有效剪枝, 更利用 FkPcore 在敏感属性上的特性(引理 2、引理 3), 最后还结合引理 4 对判定过程进行提前终止, 防止对无用点集继续进行判定, 节省了大量时间. 而方法 OptFkP 与 AdvFkP 相比, 在运行时间上的表现更优异一些. 这是因为我们提出的 TMS 算法能够有效地减少寻找各点 *P_neighbor* 时的无用操作, 提高了寻找效率, 降低了时间代价. 另外, 如图 5(d)所示, 尽管 DBpedia 数据集规模以及其中含有的点类型数和元路径数都比较大, 但是 3 种方法在这个数据集上的运行时间却不怎么高. 这是因为这个数据集过多的点类型便于借助引理 2 和引理 3 进行剪枝, 而且这个数据集中的点存在不少低度数节点, 所以这个数据集上的 3 种方法表现都比较好.



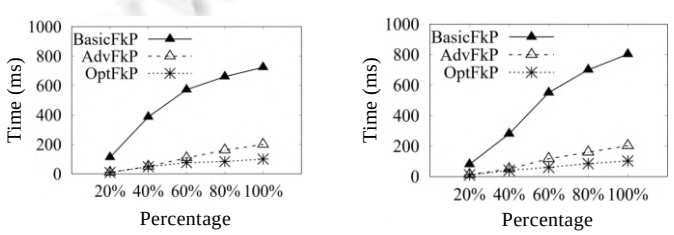
(a) Foursquare (b) DBLP

(c) IMDB (d) DBpedia

图 5 不同数据集上各方法运行时间测试

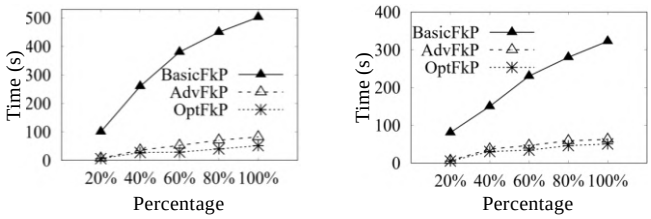
4.5.2 可扩展性测试

对表 1 中的每个数据集, 我们随机选取了 20%, 40%, 50%, 60%, 80%和 100%的点并得到了 4 个生成子图, 然后在每个子图上运行了我们实现的 3 种方法, 实验结果如图 6 所示. 结合实验结果我们可以发现, 我们所提方法在可扩展性方面表现优异.



(a) Foursquare (b) DBLP

图 6 不同数据集上各方法可扩展性测试



(c) IMDB (d) DBpedia

图 6 不同数据集上各方法可扩展性测试(续)

5 总 结

在本文中，我们研究了异质信息网络上基于属性公平的社区搜索问题. 针对这个问题，我们首先提出了基础算法 Basic-FkPcore，但是这种算法的表现并不好. 随后，我们结合本文提出的 4 个引理，对基础算法做了改进，提出了 Adv-FkPcore 算法. 最后，又结合 TMS 算法提出了优化算法 Opt-FkPcore. 我们在 4 个异质信息网络数据集上的大量实验，证明了我们所提方法的有效性和效率. 社会上由于性别、户籍差别带来的偏见屡见不鲜，我们相信，本文的研究工作势必能为克服社区搜索结果的性别或户籍偏见做出贡献.

本文提出的基于属性公平的社区搜索问题对社区中点的属性进行了严格要求，在未来，我们拟针对这一约束条件进行调整，比如给定一个阈值约束来对社区属性的公平性进行判定. 我们相信，通过这一手段，可以获得更多有趣的公平社区.

References:

[1] Sun Y, Han J. Mining heterogeneous information networks: A structural analysis approach. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20−28.

[2] Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2010. 243−252.

[3] Leroy V, Cambazoglu BB, Bonchi F. Cold start link prediction. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2010. 393−402.

[4] Huang Z, Zheng Y, Cheng R, *et al*. Meta structure: Computing relevance in large heterogeneous information networks. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 1595−1604.

[5] Fang Y, Yang Y, Zhang W, *et al*. Effective and efficient community search over large heterogeneous information networks. Proc. of the VLDB Endowment, 2020, 13(6): 854−867.

[6] Zhou H, Zhao ZY, Li C. Survey on representation learning methods oriented to heterogeneous information network. Journal of Frontiers of Computer Science and Technology, 2019, 13(7): 1082−1094 (in Chinese with English abstract).

[7] Shi C, Wang RJ, Wang X. Survey on heterogeneous information networks analysis and applications. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 598−621 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6357.htm [doi: 10.13328/j.cnki.jos.006357]

[8] Liu JW, Shi C, Yang C, Philip SY. Heterogeneous information network based recommender systems: A survey. Journal of Cyber Security, 2021, 6(5): 1−16 (in Chinese with English abstract).

[9] Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T. Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. In: Proc. of the EDBT. 2009. 565−576.

[10] Sun Y, Norick B, Han J, *et al*. PathSelClus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. ACM Trans. on Knowledge Discovery from Data (TKDD), 2013, 7(3): 1−23.

[11] Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema. In: Proc. of the KDD. 2009. 797−806.

[12] Zhou Y, Liu L. Social influence based clustering of heterogeneous information networks. In: Proc. of the KDD. 2013. 338−346.

[13] Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3-5): 75−174.

[14] Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113.

[15] Chen L, Liu C, Zhou R, *et al*. Maximum co-located community search in large scale social networks. Proc. of the VLDB Endowment, 2018, 11(10): 1233−1246.

[16] Cui W, Xiao Y, Wang H, Lu Y, Wang W. Online search of overlapping communities. In: Proc. of the SIGMOD. 2013. 277−288.

[17] Cui W, Xiao Y, Wang H, Wang W. Local search of communities in large graphs. In: Proc. of the SIGMOD. 2014. 991−1002.

[18] Fang Y, Huang X, Qin L, Zhang Y, Zhang W, Cheng R, Lin X. A survey of community search over big graphs. The VLDB Journal, 2020, 29(1): 353−392.

[19] Huang X, Cheng H, Qin L, Tian W, Yu JX. Querying *k*-truss community in large and dynamic graphs. In: Proc. of the SIGMOD. 2014. 1311−1322.

[20] Kong X, Yu PS, Ding Y, *et al*. Meta path-based collective classification in heterogeneous information networks. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. 2012. 1567−1571.

[21] Shi C, Li Y, Philip SY, *et al*. Constrained-meta-path-based ranking in heterogeneous information network. Knowledge and Information Systems, 2016, 49(2): 719−747.

[22] Jamali M, Lakshmanan L. HeteroMF: Recommendation in heterogeneous information networks using context dependent factor models. In: Proc. of the 22nd Int'l Conf. on World Wide Web. 2013. 643−654.

[23] Shi C, Zhou C, Kong X, *et al*. HeteRecom: A semantic-based recommendation system in heterogeneous networks. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2012. 1552−1555.

[24] Yu X, Ren X, Sun Y, *et al*. Recommendation in heterogeneous information networks with implicit user feedback. In: Proc. of the 7th ACM Conf. on Recommender Systems. 2013. 347−350.

[25] Verma S, Rubin J. Fairness definitions explained. In: Proc. of the Int'l Workshop on Software Fairness. 2018. 1−7.

[26] Dwork C, Hardt M, Pitassi T, *et al*. Fairness through awareness. In: Proc. of the 3rd innovations in Theoretical Computer Science Conf. 2012. 214−226.

[27] Jacobs AZ, Wallach H. Measurement and fairness. In: Proc. of the ACM Conf. on Fairness, Accountability, and Transparency. 2021. 375−385.

[28] Friedler SA, Scheidegger C, Venkatasubramanian S. On the (IM) possibility of fairness. arXiv:1609.07236, 2016.

[29] Bonald T, Massoulié L. Impact of fairness on Internet performance . In: Proc. of the ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. 2001. 82−91.

[30] Kleinberg J, Ludwig J, Mullainathan S, *et al*. Algorithmic fairness. In: Aea Papers and Proceedings, Vol.108. 2018. 22−27.

[31] Huaizhou SHI, Prasad RV, Onur E, *et al*. Fairness in wireless networks: Issues, measures and challenges. IEEE Communications Surveys & Tutorials, 2013, 16(1): 5−24.

[32] Mehrabi N, Morstatter F, Saxena N, *et al*. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 2021, 54(6): 1−35.

[33] Caton S, Haas C. Fairness in machine learning: A survey. arXiv:2010.04053, 2020.

[34] Savulescu J. Justice, fairness, and enhancement. Annals of the New York Academy of Sciences, 2006, 1093(1): 321−338.

[35] Srivastava M, Heidari H, Krause A. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2019. 2459−2468.

[36] Menon AK, Williamson RC. The cost of fairness in binary classification. In: Proc. of the Conf. on Fairness, Accountability and Transparency. PMLR, 2018. 107−118.

[37] Corbett-Davies S, Pierson E, Feller A, *et al*. Algorithmic decision making and the cost of fairness. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2017. 797−806.

[38] Sun Y, Han J, Yan X, *et al*. PathSim: Meta path-based top-*K* similarity search in heterogeneous information networks. Proc. of the VLDB Endowment, 2011, 4(11): 992−1003.

[39] Fang Y, Yang Y, Zhang W, *et al*. Effective and efficient community search over large heterogeneous information networks. Proc. of the VLDB Endowment, 2020, 13(6): 854−867.

[40] Batagelj V, Zaversnik M. An o(*m*) algorithm for cores decomposition of networks. arXiv:cs/0310049, 2003.

[41] Bonchi F, Khan A, Severini L. Distance-generalized core decomposition. In: Proc. of the ICDM. 2019. 1006−1023.

[42] Seidman SB. Network structure and minimum degree. Social Networks, 1983, 5(3): 269−287.

[43] Sozio M, Gionis A. The community-search problem and how to plan a successful cocktail party. In: Proc. of the KDD. 2010. 939−948.

[44] Zhang Z, Huang X, Xu J, Choi B, Shang Z. Keyword-centric community search. In: Proc. of the ICDE. IEEE, 2019. 422−433.

[45] Cohen J. Trusses: Cohesive subgraphs for social network analysis. Technical Report, 16(3.1), National Security Agency, 2008.

[46] Zhang Y, Yu JX. Unboundedness and efficiency of truss maintenance in evolving graphs. In: Proc. of the SIGMOD. 2019. 1024−1041.

[47] Huang X, Cheng H, Qin L, Tian W, Yu JX. Querying *k*-truss community in large and dynamic graphs. In: Proc. of the SIGMOD. 2014. 1311−1322.

[48] Yuan L, Qin L, Zhang W, *et al*. Index-based densest clique percolation community search in networks. IEEE Trans. on Knowledge and Data Engineering, 2017, 30(5): 922−935.

[49] Huang X, Lakshmanan LVS. Attribute-driven community search. Proc. of the VLDB Endowment, 2017, 10(9): 949−960.

[50] Zehlike M, Bonchi F, Castillo C, *et al*. FA*IR: A fair top-*k* ranking algorithm. In: Proc. of the ACM on Conf. on Information and Knowledge Management. 2017. 1569−1578.

[51] Serbos D, Qi SY, Mamoulis N, *et al*. Fairness in package-to-group recommendations. In: Proc. of the 26th Int'l Conf. on World Wide Web. 2017. 371−379.

[52] Beutel A, Chen JL, Doshi T, *et al*. Fairness in recommendation ranking through pairwise comparisons. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2019. 2212−2220.

[53] Pan M, Li RH, Zhang Q, *et al*. Fairness-aware maximal clique enumeration. In: Proc. of the IEEE 38th Int'l Conf. on Data Engineering (ICDE). IEEE, 2022. 259−271.

附中文参考文献:

[6] 周慧, 赵中英, 李超. 面向异质信息网络的表示学习方法研究综述. 计算机科学与探索, 2019, 13(7): 1082−1094.

[7] 石川, 王睿嘉, 王啸. 异质信息网络分析与应用综述. 软件学报, 2022, 33(2): 598−621. http://www.jos.org.cn/1000-9825/6357.htm [doi: 10.13328/j.cnki.jos.006357]

[8] 刘佳玮, 石川, 杨成, Philip SY. 基于异质信息网络的推荐系统研究综述. 信息安全学报, 2021, 6(5): 1−16.

乔连鹏(1991－), 男, 博士生, CCF 学生会员, 主要研究领域为不确定图数据管理, 稠密子图挖掘.

侯会文(1998－), 男, 硕士生, CCF 学生会员, 主要研究领域为图数据管理, 查询优化.

王国仁(1966－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为不确定数据管理, 非结构化数据管理, 分布式查询处理, 优化技术.