计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(6):45−53 [doi: 10.15888/j.cnki.csa.007931]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 ORCID 和加权跨层边聚类系数的研究者社区发现①

王毅蒙[1,2], 田　野[3], 孙善鹏[4], 周园春[1,2], 杜　一[1,2]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](中国工业互联网研究院, 北京 100102)
[4](中国科学院 软件研究所, 北京 100190)
通讯作者: 田　野, E-mail: tianye@china-aii.com

摘　要: 在开放学术的环境下, 学术交流和科研合作在学术创新与发展中发挥着重要的作用, 而帮助研究者找到合适的学术团体是促进研究者寻找科研灵感的重要途径, 现有的研究者社区发现多是着眼于科研成果的关联而忽略了研究者自身学术活动产生的关联, 因此, 本文通过分析研究者自身学术活动信息, 使用 ORCID (Open Research and Contributor ID, 开放研究者与贡献者标识) 数据构建学术信息网络, 通过综合考虑所有层次数下节点间的相似度来改进跨层边聚类系数, 提出一种基于加权跨层边聚类系数的研究者社区发现模型. 模型通过构建多种元路径抽取研究者之间的直接关联关系, 并根据不同属性关系对网络进行分层, 使用加权跨层边聚类系数计算节点间相似度, 从而将网络转化为同质网络并结合 Louvain 算法进行社区划分. 本文在人造网络和真实网络中进行实验, 根据社区实际情况对结果进行评估, 在提高了划分效果的同时避免了参数的不确定性.

关键词: ORCID; 异质网络; 社区发现; 加权跨层边聚类系数

引用格式: 王毅蒙,田野,孙善鹏,周园春,杜一.基于 ORCID 和加权跨层边聚类系数的研究者社区发现.计算机系统应用,2021,30(6):45–53. http://www.c-s-a.org.cn/1003-3254/7931.html

Community Detection of Researchers Based on ORCID and Weighted Cross-Layer Edge Clustering Coefficients

WANG Yi-Meng[1,2], TIAN Ye[3], SUN Shan-Peng[4], ZHOU Yuan-Chun[1,2], DU Yi[1,2]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](China Academy of Industrial Internet, Beijing 100102, China)
[4](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In an open academic environment, academic exchanges and scientific research cooperation are pivotal to academic innovation and development, and helping researchers find suitable academic groups contributes greatly to their inspiration. Most of the existing approaches to community detection of researchers focus on the correlation between research results and ignore that between their own academic activities. As such, this study uses Open Research and Contributor ID (ORCID) data to build a network of academic information by analyzing researchers' academic activities. The cross-layer edge clustering coefficient is improved on the basis of similarity between nodes at all levels, and then a

① 基金项目: 国家重点研发计划 (2017YFB1400203); 国家自然科学基金 (L1924075); 科技部创新方法工作专项 (2019IM020100); 北京信息科技大学高水平人才交叉培养项目 (71B2010807)
Foundation item: National Key Research and Development Plan (2017YFB1400203); National Natural Science Foundation of China (L1924075); Special Project for Innovation Method of Ministry of Science and Technology of China (2019IM020100); High-level Talent Cross-training Project of Beijing Information Science and Technology University (71B2010807)
收稿时间: 2020-10-13; 修改时间: 2020-10-27; 采用时间: 2020-11-04; csa 在线出版时间: 2021-06-01

model detecting the researcher community based on the weighted cross-layer edge clustering coefficient is proposed. The model extracts the direct correlations between researchers by constructing multiple meta-paths and stratifies the network according to different attribute relationships. Inter-node similarity is calculated with weighted cross-layer edge clustering coefficients. Then the network is transformed into a homogeneous network which is combined with the Louvain algorithm for community detection. Experiments are carried out in both artificial and real networks, and the results are evaluated according to the actual situation of the community, improving the division while avoiding the uncertainty of parameters.

Key words: ORCID; heterogeneous network; community detection; weighted cross-layer edge clustering coefficients

科研组织是学术创新的主体, 在学术创新中科研合作及学术交流发挥着越来越重要的作用, 研究者将自身的科研知识、经验和资源进行共享, 为其他研究者提供更多的灵感和思路, 创造出更多更有价值的科研成果. 因此, 挖掘出研究者之间隐含的关联关系, 寻找相关学术社区, 是值得重点关注的问题.

传统的学术社区多是着眼于研究者科技成果产生的关联进行社区发现, 忽略了研究者自身学术活动产生的关联, 如何获取并利用相关学术信息进行社区发现是本研究的重点. 随着科技信息的爆炸式增长, 不同于传统的论文数据, 科技信息数据种类更加丰富, 包括科技成果数据、科技实体数据、科技活动数据等. 在此背景下, 越来越多的学术资源网络平台应运而生, 通过科研人员唯一身份标识[1]将研究者及其学术活动信息进行关联, 如 Researcher ID[2], 帮助研究者对其出版文献进行管理, 注重对研究者著作的展示; ISNI (International Standard Name Identifier, 国际标准名称标识符)[3], 将媒体内容的贡献者赋予唯一标识, 标识相同参与者在媒体价值链上的不同身份; ORCID (Open Research and Contributor ID, 开放研究者与贡献者标识)[4], 将研究者及其学术活动精确关联, 记录研究者各项科研动态, 并与相关科研管理系统、文献数据平台、机构数据库相连接. 通过这些标识体系形成了一种底层连通的信息枢纽机制, 促进相关信息在不同系统中的流动, 可以更为便捷的得到研究者的各项学术活动及学术资源的信息[5].

因此, 本文使用 ORCID 获取研究者相关学术信息, 构建学术信息网络, 分析研究者通过不同学术活动产生的关联, 并针对网络中存在的异质性和网络层次带来的挑战, 提出一种基于加权跨层边聚类系数的社区发现模型, 挖掘出网络背后隐藏的社区结构[6], 在提高划分效果的同时对科技实体的推荐、评价、学科交叉和学科演化等相关研究均有重要意义[7].

本文余下章节中, 第 1 节对涉及到的相关工作进行概述, 第 2 节介绍基于 ORCID 的社区发现模型, 第 3 节对所提方案进行实现并对结果进行分析, 第 4 节总结全文并对未来的发展与挑战做出简要分析.

1 相关工作

如何构建学术信息网络以及如何利用学术信息进行社区发现是我们需要关注的重点.

针对学术信息网络的构建, 科研人员唯一标识符发挥了重要的作用[8], 科研人员唯一标识符能够实现对科研人员的有效标识, 提升科研成果检索效果, 便于管理科研成果和个人档案, 也可以通过对其他科研人员的信息的追踪达到寻找合作伙伴的目的, 还能将科研人员及其所属机构、参与的科研项目甚至是其他学术内容生产价值链中的潜在关联实体相链接, 从而实现科研生态系统中不同要素之间的紧密相连[9], 也可以接入相关科技领域大数据知识图谱平台[10]实现对科研数据的有效利用. ORCID, 开放研究者与贡献者标识, 以人为中心, 为全球每位研究者分配一个终生有效的唯一身份标识, 并以此为基础, 把研究者所有相关的科研活动与成果都精确地匹配并连接起来, 提高了科研人员档案的准确性. 每一位研究者 ORCID 记录中可以关联的信息包括教育经历、工作经历、发表论文、学协会会员、荣誉与奖励、大会报告、审稿贡献、科研基金等, 如图 1 所示.

图 1 中该编码采用 16 个数字表示, 每个编码分为 4 组显示, 如 0000-1234-5678-0000. 目前 ORCID 注册量已经超过 5000 000 个, 有超过 600 家学术图书馆、研究机构、资助机构和出版商会使用这些 ID 来跟踪数据, 也用于对研究者的研究成果进行追踪. 因此, 如何利用 ORCID 获取的数据进行网络的构建是我们研究的第一个重点.

针对如何利用学术信息进行社区发现, 在传统学术社区发现中大多通过分析合著网络或引文网络寻找

研究者之间的关联关系, 如图 2, 网络中包含作者、论文、会议等异质节点.

图 1 ORCID 数据内容

对于上述网络, NetClus 算法[11] 针对以论文为中心的星型学术网络, 利用排名提升聚类结果, 迭代调整每个对象的类别, 生成具有相同拓扑的输入网络的子网络合集, 每个聚类结果有相同的主题. PathSelClus 算法[12] 提出一种将元路径与聚类相结合的算法, 通过预先为每个聚类提供一部分种子节点, 系统学习到元路径的权重, 根据权重产社区, 叠加不同元路径的聚类结果生产最终社区. Lu 等提出了 Hete_MESE 多维社区检测算法[13], 首先将异构信息网络中的多个实体类型之一指定为社区中心节点类型, 并相应地提取复用网络, 然后, 基于复用网络检测重叠的节点中心社区, 将其视为种子社区, 吸收其他实体类型以利用种子扩展产生异质社区. 文献 [14] 基于 Salton 方法计算作者间相似度以评估合著关系强弱, 将节点间的边作为聚类对象, 采用凝聚式层次聚类进行学术社区发现. 文献 [15] 以直接引用关系构建显性关联, 以引文抽取出的兴趣标签构建隐性关联, 用以衡量研究者之间关系的强弱从而进行社区发现. 而面对大规模的学术信息网络, 如图 3, 网络中节点种类更多, 关联关系更复杂, 复杂的网络结构对社区发现带来了新的挑战.

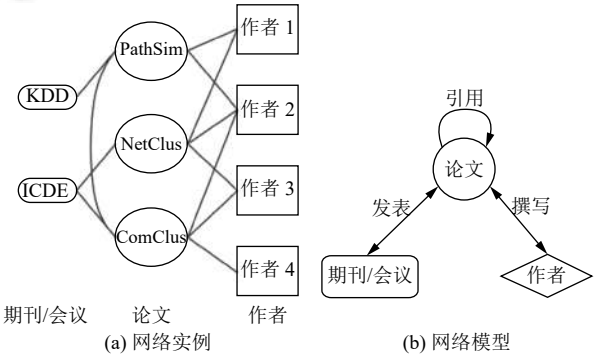

图 2 合著网络示例

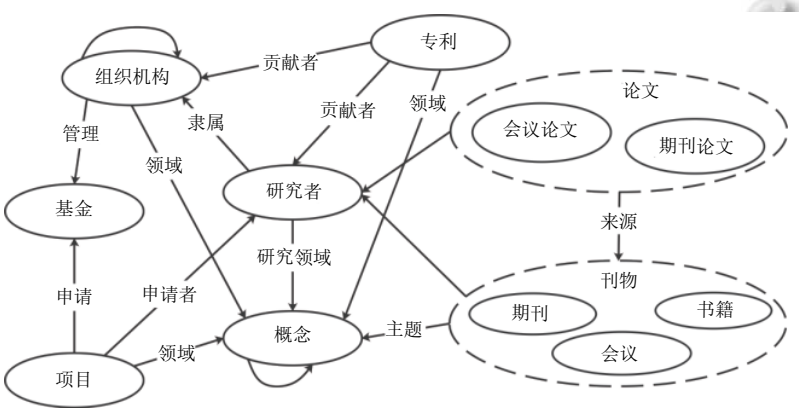

图 3 复杂学术信息网络

针对异质网络的社区发现研究已得到了学者的广泛关注, 本文重点阐述多层网络社区发现的相关研究成果. 文献 [16] 采用多目标方法, 在第一层应用经典社区发现算法, 对其余的连续层, 采用最大化当前层模块度和前一层划分的社区结构的相似性双目标优化方法来发现社区. 文献 [17] 对每一层网络应用经典社区发现算法并用集成聚类方法合并划分的社区来发现社区. 文献 [18] 提出了一种基于元路径嵌入的聚类方法 MPEClus, 将原始网络转换为具有由元路径指定的拥有独立语义的多个子网, 使用近似通勤嵌入学习节点的向量表示, 并针对不同度量空间中学习的节点向量进行社区发现. 文献 [19] 使用基于频谱聚类和低秩矩阵分解的方法组

合多层网络的多层信息来进行社区发现. 文献 [20] 通过使用跨层边聚系数计算节点间相似度并通过不断更新损失函数实现多层网络社区划分. 因此, 如何解决学术信息网络中异质性和网络层次带来的挑战, 从而进行社区发现, 也是我们需要研究的重点.

2 基于 ORCID 和加权跨层边聚类系数的社区发现模型

本文基于 ORCID 获取的数据集, 分析研究者及其学术活动信息构建学术信息网络, 寻找研究者之间多属性的关联关系并计算研究者之间的相似度, 从而进行学术社区的发现, 本文算法流程图如图 4 所示.

2.1 构建 ORCID 异质网络

通过分析 ORCID 数据中包含的学术活动信息, 可以发现研究者之间通过不同学术信息可以产生多种关联, 将不同学术信息作为不同类型节点从而构建异质网络, 网络中包含研究者节点 P, 教育经历节点 E、工作经历节点 W、受邀职位节点 I、服务单位节点 S、学术领域节点 D, 如图 5 所示, 同时, 不同节点之间也存在不同类型的关联关系. 通过 ORCID 异质网络, 不仅可以快速获取研究者相关的学术活动信息, 也可以通过某些学术活动查询到相关联的研究者, 不同研究者通过中间学术活动节点也可以取得不同属性的关联.

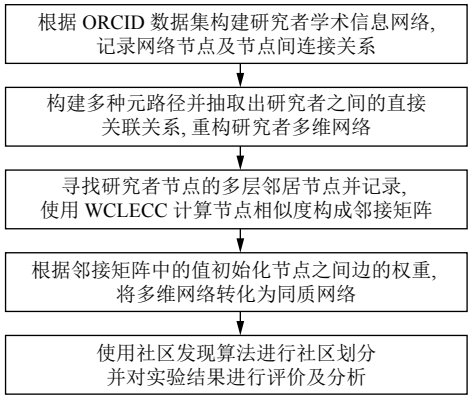

图 4 基于 ORCID 的学术社区发现算法流程

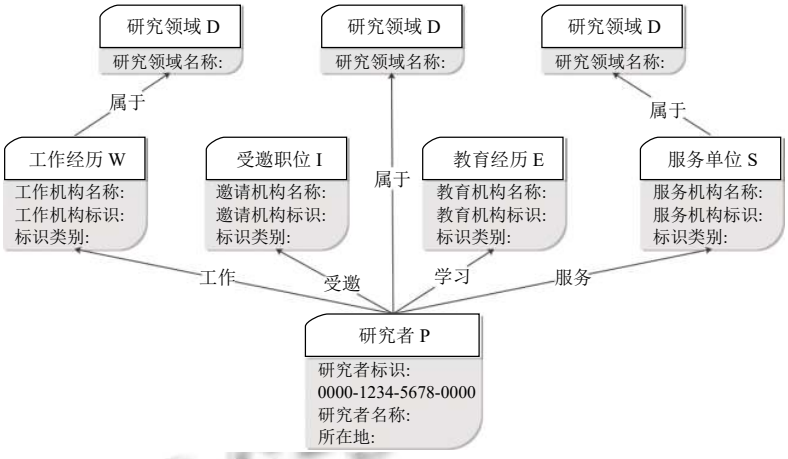

图 5 ORCID 异质网络

2.2 根据元路径抽取研究者多维异质网络

由于构建的 ORCID 异质网络中存在大量的学术活动节点, 研究者之间并非直接相连, 而是存在不同的路径. 不同路径连接的研究者之间存在不同语义的关联关系, 构成了多种元路径, 图 6 展示了 ORCID 异质网络中存在的部分元路径, P 为研究者节点、D 为研究领域节点、W 为工作单位节点, P3D2P4 表示 P3 和 P4 有相同的研究领域, P1W1P2 表示 P1 和 P2 在相同的单位工作过, P1W1D1W2P3 表示 P1 和 P3 有相同领域内的工作经历.

多种元路径的存在既无法直观发现研究者节点之间的关联关系, 也增加了计算研究者相似度的难度. 因此, 本文通过不同元路径提取出研究者节点之间的多种直接关联关系, 从而构成研究者多维异质网络, 网络中仅包含研究者节点一种节点和多种不同属性的边. 元路径选择如表 1 所示, 从而根据新的关联关系重构研究者多维异质网络, 解决了 ORCID 异质网络中节点多样性而产生的社区划分问题.

2.3 节点相似度计算

基于研究者多维异质网络, 本文综合考虑研究者节点间的多种属性关联关系来计算多维网络中节点间的相似度. 本文考虑使用 Brodka 等提出的跨层边聚类

系数 *CLECC*[20] 可以用来计算多维网络中节点间的相似度, 但是在计算过程中, 只能针对某一层次计算节点间相似度, 通过多次尝试选出最优结果, 可控性不足, 尤其是在网络层数较大的情况下, 计算开销和存储开销很大. 因此, 本文提出加权跨层边聚类系数 *WCLECC*, 解决层次数不可控的问题, 综合考虑层次数的所有可能值, 对于在所有层次数下取得的相似度值进行权重处理, 层次数越高权重越大, 对计算相似度的影响越大. 加权跨层边聚类系数 *WCLECC* 计算公式如下:

$$WCLECC(x,y)=\frac{1}{z}\sum_{a=1}^{|L|}\frac{2a}{|L|(|L|+1)}\cdot\frac{|MN(x,a)\cap MN(y,a)|}{|(MN(x,a)\cup MN(y,a))/\{x,y\}|} \quad (1)$$

其中, $|L|$为最大网络层数. $MN(x,a)$ 为 x 节点的多层邻居集合, 是指与节点 x 有 a 层或 a 层以上关联的邻居节点的集合, z 为归一化因子. 以此做为衡量节点间紧密度的指标, 充分考虑了网络中不同层的稀疏程度, 且不需要进行参数的调整, 可以更准确的衡量节点间的关系强度.

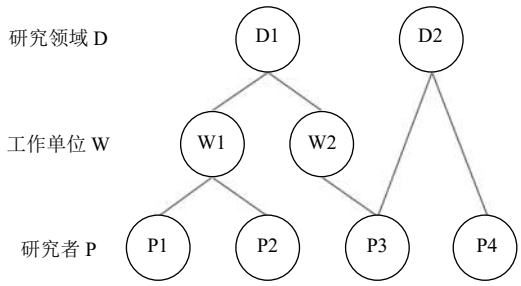

图 6 ORCID 异质网络中的元路径

表 1 ORCID 网络元路径语义表

编号	元路径	路径语义	新的关联关系
P1	PDP	研究者有共同的研究领域	r_d
P2	PEP	研究者在相同的学校学习过	r_e
P3	PWP	研究者在相同单位工作过	r_w
P4	PWDWP	研究者有相同领域的工作经历	
P5	PSP	研究者有相同的服务单位	r_s
P6	PSDSP	研究者有相同领域的服务经历	
P7	PIP	研究者有相同职位邀请的经历	r_i

2.4 社区发现

通过使用 *WCLECC* 作为衡量节点间的相似度指标, 将多维网络转化为同质网络, 然后运用社区发现算法进行社区划分. 将节点 i 加入到节点 j 所在社区产生的模块度增量如式 (2):

$$\Delta Q=\left[\frac{\Sigma_{in}+k_{i,in}}{2m}-\left(\frac{\Sigma_{tot}+k_i}{2m}\right)^2\right]-\left[\frac{\Sigma_{in}}{2m}-\left(\frac{\Sigma_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \quad (2)$$

$\sum_{in}$ 表示社区内边的权重之和, $\sum_{tot}$ 表示与社区内节点相连的边的权重之和, $k_{i,in}$ 表示社区内节点与节点 i 的边权重之和. 算法流程如下所示:

1) 构建网络节点邻接矩阵 A, 且将值均置为 null;

2) 遍历网络中的每一个节点 x, 并记录该节点的所有邻居节点 $Y\{y:y\in MN(x)\}$;

3) 计算每一对节点 (x,y) 的相似度 $WCLECC(x,y)$, 并更新邻接矩阵 $A(x,y)$ 的值;

4) 在邻接矩阵 A 中, 当 $A(x,y)$!=null, 在新的网络中连接 x 节点与 y 节点并将 $WCLECC(x,y)$ 作为边的权重, 重构研究者同质网络 G';

5) 将 G'中每个节点作为一个单独的社区, 社区数与节点数相同;

6) 对 G'每一个节点 x, 依次将 x 加入其邻居所在社区之中, 计算加入前后的模块度变化情况 ΔQ, 记录 ΔQ 最大的邻居节点 n, 如果 max$\Delta Q>0$, 则把节点 x 加入到 n 所在社区, 否则不改变 x 所在社区;

7) 重复步骤 6), 直到所有节点所属社区不再变化;

8) 对产生的社区进行压缩, 将每一个社区看作一个新的节点, 社区内边的权重之和当作社区自身环的权重, 社区间边的权重之和当作新节点之间边的权重;

9) 重复步骤 5), 直到全图模块度不再发生变化;

10) 选出模块度最大时网络的社区划分结果, 即为最终社区划分情况.

3 实验与分析

3.1 社区评价标准

常用的评价无监督社区划分结果优劣的指标为模块度 (modularity)[21]. 其物理意义是社区内节点的连边数所占的比例与随机放置情况下社区内节点期望连边数的比例的差值, 定义如下:

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j) \quad (3)$$

$$\delta(u,v)=\begin{cases}1, & \text{when } u=v\\ 0, & \text{else}\end{cases} \quad (4)$$

其中, A_{ij} 是节点 i 和节点 j 之间边的权重, k_i 为所有与节点 i 相连的边的权重之和, C_i 为节点 i 所属的社区,

m 为图中所有边的权重之和. 通常取值范围在 [−1/2, 1] 之间, 其值越靠近 1, 表明网络划分结果越好.

3.2 实验结果及分析

3.2.1 ORCID 学术信息网络和研究者多维异质网络构建结果

本文通过对 ORCID 数据集中研究者、教育经历、工作经历、受邀职位、服务单位的数据量进行统计, 如图 7 所示.

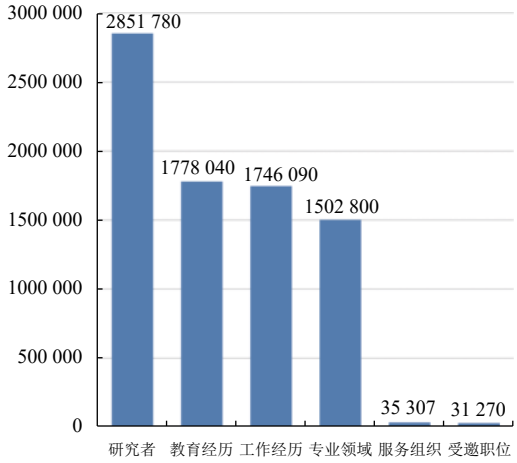

图 7 ORCID 不同属性数据量统计

本文样本的选择根据 ORCID 标识符的 11 种尾号 (0~9、X) 分层选取, 每种尾号的数据选取 1 万条, 并去除掉未包含任何属性信息的数据, 共选取 3 组样本, 每组 10 万余名研究者的信息进行实验, 构建 ORCID 学术信息网络, 网络具体数据如表 2 和表 3 所示.

表 2 ORCID 学术信息网络各节点数量统计

样本编号	研究者节点P	教育经历节点E	工作经历节点W	受邀职位节点I	服务单位节点S	专业领域节点D
1	103 816	36 528	63 633	10 700	11 461	29 451
2	103 832	36 605	63 792	10 655	11 422	29 406
3	103 837	37 081	63 677	10 696	11 425	29 503

表 3 ORCID 学术信息网络各属性边数量统计

样本编号	P-E	P-W	W-D	P-I	P-S	S-D	P-D
1	158 920	155 377	127 915	155 568	173 435	75 029	37 434
2	158 815	155 175	128 524	156 129	171 985	75 242	37 657
3	159 755	155 650	128 962	156 360	174 007	75 060	37 421

在构建好的 ORCID 异质网络中, 通过表 1 中的元路径抽取研究者节点间不同属性的直接关联关系, 构建研究者多维异质网络, 网络中只含有研究者节点及不同属性连边, 网络具体数据如表 4 所示.

通过元路径的抽取, 可以将 ORCID 异质网络中多种类多属性的节点和边简化为只存在研究者节点及其之间多属性边的多维网络, 减少了网络节点类型, 避免了其余组织机构节点对社区划分产生的影响, 降低了网络的复杂性和计算的复杂性.

表 4 研究者多维异质网络连边数量统计

样本编号	学习R_E	工作R_W	服务R_S	受邀R_I	专业R_D
1	5772 469	4895 584	5025 026	3152 662	32 994
2	5630 276	5246 163	4397 565	3121 129	34 374
3	5744 093	5284 086	4414 123	3245 271	34 514

3.2.2 社区划分结果

(1) 通过构建人造稀疏网络和稠密网络对本文算法进行实验, 测试 *WCLECC* 与 *CLECC* 在取不同 a 值的情况下对网络的划分取得的效果, 以此检测是否通过 *WCLECC* 避免了 *CLECC* 参数的不确定性对实验产生的影响且能取得优于 *CLECC* 的实验结果.

① 图 8 为 4 层稀疏网络中每层的初始连边情况.

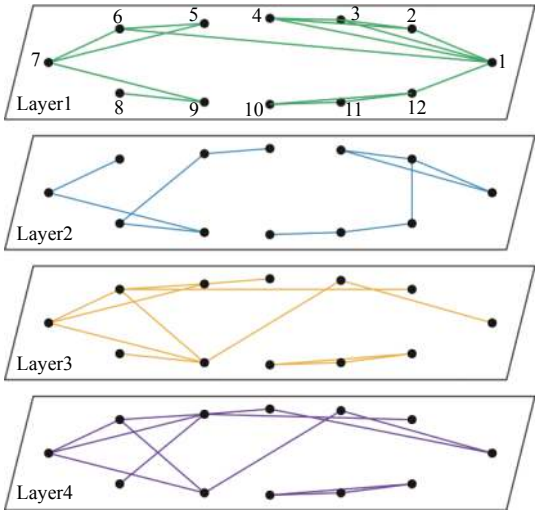

图 8 稀疏网络各层初始情况

实验结果如表 5 所示, 可以看出使用 *CLECC* 在 a=2 时, 网络划分可以取得最大模块度 Q, 而使用 *WCLECC* 划分的社区数量和成员与其相同且模块度提高了 1.85%, 实验效果更好.

② 图 9 为 4 层稠密网络中每层的初始连边情况.

实验结果如表 6 所示, 可以看出使用 *CLECC* 在 a=3 时, 网络划分可以取得最大模块度 Q, 而使用 *WCLECC* 划分的社区数量和成员与其相同且模块度提高了 1.65%, 实验效果更好.

通过上述实验可知, 使用 *CLECC* 进行社区划分时, 在稀疏网络中 a 取较小值可以得到更优的实验结

果, 在稠密网络中 a 取较大值可以得到更优的实验结果. 究其原因, 当网络稀疏时, 高层次邻居节点远少于低层次邻居节点, 当 a 取较大值时会造成部分节点间相似度丢失, 影响社区划分的准确性, a 取较小值时会有更多的邻居节点参与相似度的计算, 提高计算准确性. 而当网络稠密时, 高层次邻居节点与低层次邻居节点数量相近, a 取较大值能更准确计算出节点间的相似度, 使网络划分更准确. 针对稀疏程度不确定的网络, 使用 *CLECC* 进行社区划分必须要依次尝试 a 取值的所有可能值才能找到最优的实验结果, 而 *WCLECC* 针对 *CLECC* 参数不确定的问题, 综合考虑了 a 参数的所有可能取值, 简化了参数选择的过程, 并且在取得相同划分结果的同时能取得更优的实验结果. 因此, 当网络稀疏程度明确时, 可以考虑使用 *CLECC* 进行计算, 也可以使用 *WCLECC* 进行计算, 当网络稀疏程度不明确时, 为避免多次尝试不同参数可以使用 *WCLECC* 进行计算从而进行社区划分.

表 5 稀疏网络社区划分结果表

度量指标	社区数量	社区成员	模块度
CLECC (a=1)	4	社区1: (1,2,3,4,5) 社区2: (6,7,9) 社区3: (8) 社区4: (10,11,12)	0.606
CLECC (a=2)	4	社区1: (1,2,3,4) 社区2: (5,6,7,9) 社区3: (8) 社区4: (10,11,12)	0.702
CLECC (a=3)	7	社区1: (1) 社区2: (2) 社区3: (3) 社区4: (4) 社区5: (5,6,7,9) 社区6: (8) 社区7: (10,11,12)	0.576
CLECC (a=4)	3	社区1: (1,2,3,4,5) 社区2: (6,8,7,9) 社区3: (10,11,12)	0.631
WCLECC	4	社区1: (1,2,3,4) 社区2: (5,6,7,9) 社区3: (8) 社区4: (10,11,12)	0.715

同时, *WCLECC* 对于 *CLECC* 的改进主要在于参数选择的优化, 针对稀疏程度不明的网络可以减少对不同参数的尝试并能得到更优的结果, 但 *WCLECC* 需要同时考虑各个层次的邻居, 增加了部分计算时间, 但整体时间仍保持在同样的量级, 对时间开销方面并未造成过大的影响.

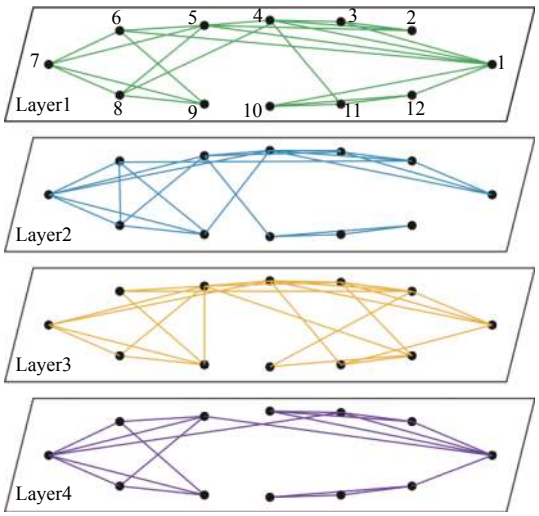

图 9 稠密网络各层初始情况

表 6 稠密网络社区划分结果表

度量指标	社区数量	社区成员	模块度
CLECC (a=1)	2	社区1: (1,2,10,11,12) 社区2: (3,4,5,6,7,8,9)	0.311
CLECC (a=2)	3	社区1: (1,2,3,4) 社区2: (5,6,7,8,9) 社区3: (10,11,12)	0.539
CLECC (a=3)	3	社区1: (1,2,3,4) 社区2: (5,6,7,8,9) 社区3: (10,11,12)	0.726
CLECC (a=4)	3	社区1: (1,2,3,4) 社区2: (5,6,8,7,9) 社区3: (10,11,12)	0.557
WCLECC	3	社区1: (1,2,3,4) 社区2: (5,6,7,8,9) 社区3: (10,11,12)	0.738

(2) 在构建好的研究者多维异质网络中运行本文算法进行社区发现. 图 10 为 3 次实验过程中社区划分中模块度随迭代次数的变化, 选取模块度最高时的划分结果作为最终的实验结果. 表 7 为 3 次实验中划分的社团数和模块度结果的对比.

由上述结果可以看到, $a \geq 3$时结果产生了突变, 模块度的值大幅提高同时划分的社区数量过多, 可能产生了大量孤立节点和小成员数的社区, 无法满足社区发现的目的. 针对上述情况, 本文对所划分的社区进行了分析, 统计所划分社区中孤立节点社区的占比情况和拥有不同成员数的社区占比情况. 图 11 为 3 次实验中未被划分进社区的孤立节点数占总节点数的比例情况.

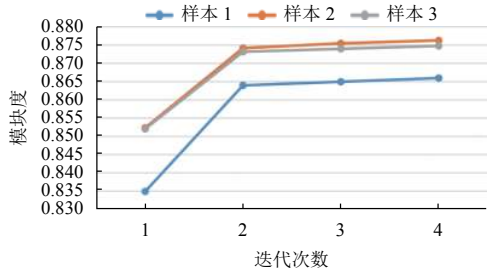

图 10 社区划分中模块度随迭代次数的变化

表 7 社区划分结果

样本编号		样本1	样本2	样本3
研究者节点数		103 816	103 832	103 837
社区数	*WCLECC*	368	388	371
	CLECC(*a*=1)	181	182	177
	CLECC(*a*=2)	3941	5458	5431
	CLECC(*a*=3)	62 168	85 404	85 525
	CLECC(*a*=4)	102 868	103 554	103 543
	CLECC(*a*=5)	103 594	103 558	103 545
模块度	*WCLECC*	0.866	0.876	0.875
	CLECC(*a*=1)	0.784	0.787	0.788
	CLECC(*a*=2)	0.867	0.865	0.864
	CLECC(*a*=3)	0.999	0.999	0.999
	CLECC(*a*=4)	0.997	0.994	0.995
	CLECC(*a*=5)	0.990	0.994	0.995

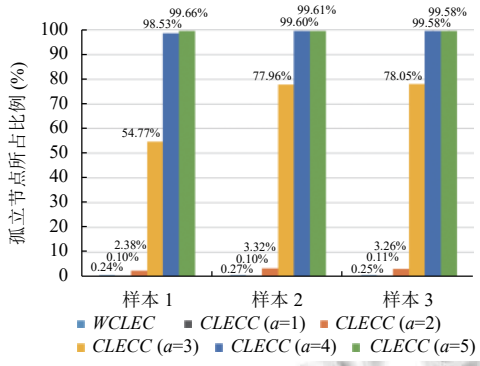

图 11 孤立节点占比情况

图 12 展示了 3 次实验中社区成员数超过不同阈值的社区占比情况.

通过观察上述结果, 当 a=1 时, 实验结果中模块度的值最低, 划分社区数最少, 虽然所划分的社区能覆盖最多的节点, 但整体来看划分效果不佳; 当 a=2 时, 能取得较好的模块度结果及适中的社区数, 孤立节点占比较低, 虽然成员数超过不同阈值的社区数量较少, 产生了大量的小社区团体, 但整体来看取得了较好的实验结果; 当 a>2 时, 虽然模块度的值均能接近理论最优值, 但划分的社区数量过多, a=3 时, 孤立节点占比超过 50%, 且社区成员超过 10 人的社区比例仅在样本 1 中达到 20% 以上, 其余均低于 10%, 当 a>3 时, 所划分社区几乎全是孤立节点社区, 未起到社区划分的真正意义, 实验结果不佳. 究其原因, 由于不同层次的网络稀疏程度不同, 当层数越深, 节点间有多层关联的邻居越少, 仅有少量节点间拥有多层次的关联关系, 忽略了低层次关联产生的影响. *WCLECC* 很好的解决了这一问题, 充分考虑了所有层的关联关系, 模块度的值和孤立节点的占比情况均优于 a=2 的结果, 在成员数符合阈值的社区比例中也能取得最优的结果, 可见使用 *WCLECC* 减少了孤立节点和小成员数社区的产生, 整体来看取得的效果最佳.

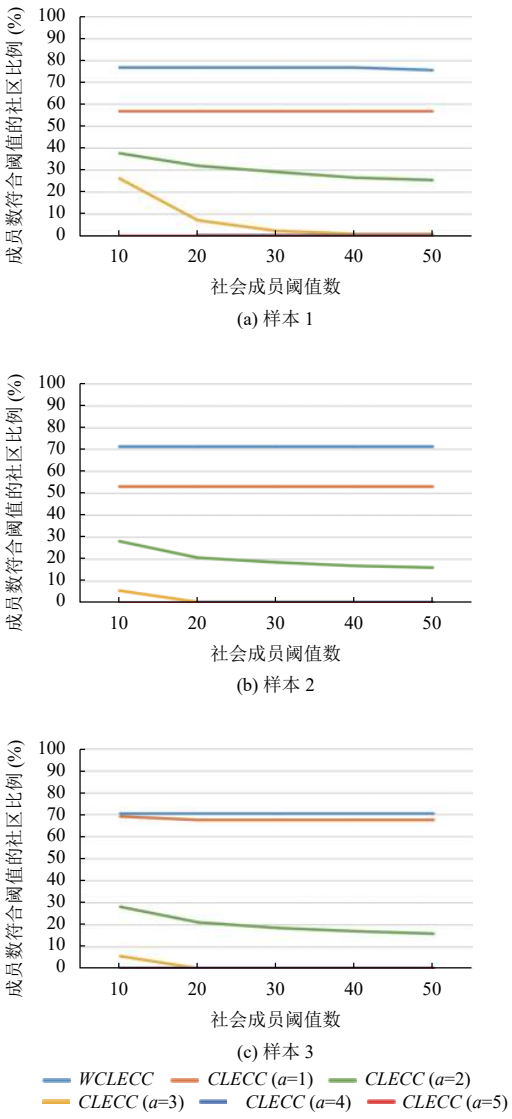

图 12 成员数符合阈值的社区占比情况

综上所述, 通过使用研究者学术活动信息构建 ORCID 异质网络, 并使用 *WCLECC* 能取得最优的社

区划分结果, 既充分考虑了研究者节点间的多层关联关系, 又避免了参数的不可控, 同时产生的社区覆盖了较多的研究者节点, 减少了孤立节点的出现, 也减少了小成员数社区的出现, 划分出了高质量的社区, 得到了较好的实验结果.

4 结语

本文通过对 ORCID 数据进行分析, 使用研究者学术活动构建科研信息网络进行学术社区的发现, 通过元路径抽取出研究者节点间的直接关联关系, 降低了异质网络的复杂度, 避免了中间节点对社区划分产生的影响, 提出加权跨层边聚类系数解决了多层网络中节点相似度的度量问题, 改善了跨层边聚类系数的参数不可控性, 充分利用研究者的学术信息去寻找其学术团体, 对学术社区发现提出了一种新的思路. 在人造网络和真实数据集上进行实验, 均取得了较好的实验结果. 同时, 本文还存在一定的问题, 如尚未对全部数据进行实验, 不同属性信息对划分结果的影响等也值得更进一步的考虑, 后续的工作将针对这些问题进行进一步的研究.

参考文献

1 郑琳. 面向机构的科研人员唯一标识符应用与发展研究. 图书与情报, 2018, (1): 134–140. [doi: 10.11968/tsyqb.1003-6938.2018017]

2 Researcher ID. http://www.researcherid.com/#rid-for-researchers.

3 Adamich T. Linked data identifiers: Part 1 –International Standard Name Identifier (ISNI). Technicalities, 2014, 34(1).

4 刘润达, 王运红. 开放研究人员及贡献者唯一标识 (ORCID) 概述. 情报科学, 2013, 31(11): 86–90. [doi: 10.13833/j.cnki.is.2013.11.027]

5 Akers KG, Sarkozy A, Wu W, *et al*. ORCID author identifiers: A primer for librarians. Medical Reference Services Quarterly, 2016, 35(2): 135 –144. [doi: 10.1080/02763869.2016.1152139]

6 赵卫绩, 张凤斌, 刘井莲. 复杂网络社区发现研究进展. 计算机科学, 2020, 47(2): 10–20. [doi: 10.11896/jsjkx.190100214]

7 周园春, 王卫军, 乔子越, 等. 科技大数据知识图谱构建方法及应用研究综述. 中国科学: 信息科学, 2020, 50(7): 957–987. [doi: 10.1360/SSI-2019-0271]

8 黄国彬, 郑琳. 科研人员唯一标识符的组成与应用研究. 图书情报工作, 2015, 59(4): 25–31. [doi: 10.13266/j.issn.0252-3116.2015.04.004]

9 Aerts R. Digital identifiers could keep up with authors' moves. Nature, 2008, 454(7204): 575. [doi: 10.1038/454575c]

10 周园春, 常青玲, 杜一. SKS: 一种科技领域大数据知识图谱平台. 数据与计算发展前沿, 2019, 1(1): 82–93. [doi: 10.11871/jfdc.issn.2096-742X.2019.01.009]

11 Sun YZ, Yu YT, Han JW. Ranking-based clustering of heterogeneous information networks with star network schema. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France. 2009. 797–806. [doi: 10.1145/1557019.1557107]

12 Sun YZ, Norick B, Han JW, *et al*. PathSelClus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. ACM Transactions on Knowledge Discovery from Data, 2013, 7(3): 11. [doi: 10.1145/2500492]

13 Lu ML, Qu ZH, Wang ZH, *et al*. Hete_MESE: Multi-dimensional community detection algorithm based on multiplex network extraction and seed expansion for heterogeneous information networks. IEEE Access, 2018, 6: 73965–73983. [doi: 10.1109/ACCESS.2018.2883638]

14 谷瑞军, 陈圣磊, 陈耿, 等. 复杂合著网络中的重叠社团发现与可视化. 图书情报工作, 2012, 56(12): 72–76, 59.

15 逯万辉, 谭宗颖. 基于显性-隐性二元关系的学术社区发现方法研究. 情报科学, 2018, 36(1): 130–134. [doi: 10.13833/j.issn.1007-7634.2018.01.023]

16 Amelio A, Pizzuti C. Community detection in multi-dimensional networks. 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. Limassol, Cyprus. 2014. 352–359.

17 Kanawati R. Community detection in social networks: The power of ensemble methods. 2014 International Conference on Data Science and Advanced Analytics. Shanghai, China. 2014. 46–52.

18 Zhang JY, Yang XP, Wang L. Clustering via meta-path embedding for heterogeneous information networks. 2020 IEEE International Conference on Knowledge Graph. Nanjing, China. 2020. 188–194. [doi: 10.1109/ICBK50248.2020.00036]

19 Paul S, Chen YG. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. The Annals of Statistics, 2020, 48(1): 230–250.

20 Bródka P, Filipowski T, Kazienko P. An introduction to community detection in multi-layered social network. Proceedings of the 4th International Conference on Information Systems, E-Learning, and Knowledge Management Research. Mykonos, Greece. 2013.185–190.

21 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113. [doi: 10.1103/physreve.69.026113]