

基于网络结构变化测度科研论文重要度的方法研究^{*}

郭红梅 陈星宇 李超 马峥
(中国科学技术信息研究所, 北京 100038)

摘要: 知识的传承和交流蕴涵在学者的交互过程中, 形成复杂的动态网络。从网络演化动力学视角出发, 通过探索某节点对原有网络结构的改变来测度该节点的重要度, 并将该思路沿用到文献网络中以测度节点论文的重要度。基于引用关系构建作者引用网络和主题引用网络, 将每篇待测论文网络与学术社区原有基础网络叠加融合后, 以节点数、节点度、边数、边权重、网络密度等多个指标测度每篇论文加入前后网络结构的改变, 利用网络结构指标的综合得分衡量节点论文重要度。选用2013—2018年Web of Science数据库中传染病领域数据为实验数据, 以2013—2017年作者引用网络和主题引用网络为学术社区基础网络, 计算2018年每篇论文引文网络加入后对原有网络结构改变的综合得分作为节点论文的重要度, 再以2018年引文数据为参照数据, 证明论文重要度得分与被引频次有较强的相关性。

关键词: 重要度; 学术社区; 作者引用网络; 主题引用网络

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2023.11.007

引文格式: 郭红梅, 陈星宇, 李超, 等. 基于网络结构变化测度科研论文重要度的方法研究[J]. 数字图书馆论坛, 2023 (11): 53-61.

科研论文是科学技术发展过程中重要的知识载体, 记录着人类发现、创造的科技成果和知识信息, 推动着科学的快速进步和学术交流社区的不断扩展。每篇科研论文并不是孤立存在的个体, 而是学者们在继承原有知识结构基础上, 经过不断的探索改进产生的新知识和新发现。科研论文作者的交互承载着知识的流动和传承, 论文作者是交互过程的主体, 论文所蕴含的知识是交互的客体, 论文之间的相互引用是最主要的交互形式, 在此过程中形成了错综复杂的动态学术社区网络, 每篇论文都是学术社区网络中的重要节点^[1], 网络中节点的作用和地位随着网络结构的改变而变化。本文从网络演化动力学视角出发, 探索通过某节点对原有社区网络结构的改变程度来测度该节点的重

要度, 并将该思路沿用到文献网络中以测度节点论文的重要度。

1 相关研究进展

面对海量的科研产出, 如何快速识别和遴选出不同领域重要的科研论文是科研评价、检索和推荐服务面临的重要问题。科研论文重要度评价的本质是学术价值和影响力的判断, 学者们从文献计量指标、文本内容特征和社会网络特征等多种维度进行了探索, 提出了新颖性、影响力、创新性、原创性、价值性等多个测度指标, 这些指标从不同维度对论文重要性进行了揭示。

收稿日期: 2023-08-14

^{*}本研究得到中国科学技术信息研究所创新研究基金面上项目“面向高端交流平台建设的文献多维属性聚类 and 代表作成果遴选方法”(编号: MS2023-07)资助。

1.1 基于文献计量的相关评价方法研究

文献计量指标是最基础的评价方法,其中以引文为核心的指标占据重要地位,包括被引频次、H指数、G指数、Z指数等以及这些指数的改进方法^[2]。学者们基于Google Scholar、Scopus、Web of Science等不同引文数据库探索,对论文、期刊、作者等不同对象进行重要度评价。Tahamtan等^[3]总结了可能影响论文被引频次的相关因素。Boyack等^[4]提出从期刊的重要性、参考文献的重要性和作者名誉来预测论文的重要性。耿树青等^[5]探索了基于引用情感的论文学术影响力评价方法。随着社交平台发展,Altmetrics指标应用日益广泛,包括下载数量、社交网站的共享数量、点赞和订阅数等^[6]。Sathianathan等^[7]根据Altmetrics指标与论文被引频次,通过构建多元回归模型得到综合影响力指标SoME_Impact。郭倩影等^[8]对Altmetrics数据源的选择、评价指标的设计、评价策略和结果评估方法进行了深入总结。还有学者尝试通过建立多维度的评价体系对学术成果影响力展开综合性的评价^[9]。

1.2 基于文本内容特征的相关评价方法研究

随着文本挖掘技术的快速发展,学者借助自然语言处理和智能化技术深入文本内容层次,基于研究主题识别和评价科研论文的重要性。Shibayama等^[10]提出了词嵌入的论文创新性评价方法,认为科研论文的价值体现在它被其他科学家重用和继承的过程中。Wang等^[11]提出从论文微观层次测度特定知识元素创新性的方法,不仅确保论文发表前的质量,还可辅助学者快速识别论文中有关知识内容。He等^[12]提出通过构建词嵌入模型,跟踪术语的语义变化计算余弦相似度来比较语义属性,并衡量特定时期内科学界研究主题的新颖性。Luo等^[13]提出了基于问题-方法组合的科技论文新颖性度量方法。Chen等^[14]构建科研论文创新观点自动抽取模型。曹树金等^[15]基于深度学习模型、语义角色信息探索创新段定位方法,通过BERT预训练模型识别出全文中的创新句,通过语义角色信息匹配创新段落。Mishra等^[16]提出基于论文中医学主题词的时间分布规模来衡量生物医学领域论文的新颖性。贺婉莹^[17]探索了基于机器学习的论文学术创新力评价方法。

1.3 基于网络特征的相关评价方法研究

随着文献数量的指数级增长,研究内容随着知识流动形成一个关联网,而网络的拓扑结构对网络的演化发展存在影响,通常认为连接点最多的点是关键参与者,参与最大规模的传播过程^[18]。在社会网络理论中,节点对传播的重要度通常与中间中心性有关,中间中心性衡量的是有多少最短路径穿过该节点,说明此节点在整个社会网络中的影响力和重要度。图或网络中的节点和边可以清晰反映对象及对象之间的关系,随着图或网络挖掘方法的提出,很多学者尝试利用网络结构特征来测度节点和边的重要度。Kitsak等^[19]提出K-壳分解法来识别网络中最有效的“传播者”。Zhang等^[20]提出了基于引文相关性和作者贡献的加权网络,通过改进异构文献计量网络和W-Rank算法,量化论文影响力。Min等^[21]通过计算平均聚类系数、网络密度、关联性,证实了科学新颖性与引用模式相互作用。Amplayo等^[22]提出了一种基于网络的特征提取方法,构建了作者和文档的关联网以及以关键词、主题词为节点的关联网,论文的新颖性由网络内节点数量随时间的增加和减少来判断。Amplayo等^[23]基于网络思想对比了文本向量、TF-IDF、SVM等多种文本新颖性计算方法的有效性。谭琳洁等^[24]对LeaderRank算法进行改进,并兼顾节点新鲜度、邻近节点数量和质量,构建了基于含时引文网络的论文现时影响力评价模型。

综上,科研论文内容结构复杂,学术生态网络庞大,只有综合考虑多个维度才能对论文进行全面客观的评价。定量评价总体属于形式范畴,没有深入论文内容层面,容易导致过度强调被引频次等外在量化指标。科研论文是由一系列具有语义内容的句子按照逻辑结构组合而成的,蕴涵着丰富的语义知识和主题信息,由于语言的复杂性和多义性,基于文本内容特征的评价模型易受词汇和语义的影响。基于节点度、中心性、特征向量等指标的评价方法不能反映节点在网络全局范围内的作用。论文所蕴含知识的流动和传承是在学术社区内学者们思想交互过程中实现的,交互过程中形成了多种复杂的网络,知识流动和传承的效果体现为对网络结构的改变程度,尤其是体现在对网络结构的优化或完善的过程中。因此,本文将每篇科研论文置于学术社区网络,假设论文对已有学术社区网络结构的改变在一定程度上可反映论文的重要度和潜在影响力,进而探索通过计算每篇论文加入前后学术社区网络结构的

改变程度来测度节点论文重要度。

2 基于网络结构变化的节点重要度测度思路和方法

以往基于网络测度节点重要度时,通常将所有节点置于同一个静态网络中进行对比,但是网络是动态变化的,尤其是体现知识流动的引文网络。随着学术社区逐步壮大,知识网络结构越来越复杂,仅基于某一时刻的网络并不能全面衡量节点重要度。因此,提

出通过分析在每个节点加入后原有网络结构的改变来衡量该节点的重要度,网络结构的改变通常体现为节点数和边数的变化、节点度和边权重的变化以及网络密度等指标的变化,具体指标为新增节点数、新增节点度、新增边数、新增边权重以及网络密度变化,通过综合多个指标测度节点重要度,具体思路方法见图1。其中:原有网络中的不同P点表示存在关联关系的不同论文;新网络中的A点表示新加入的网络节点,表示新增的待测节点,各个P点为与A点存在关联的节点。

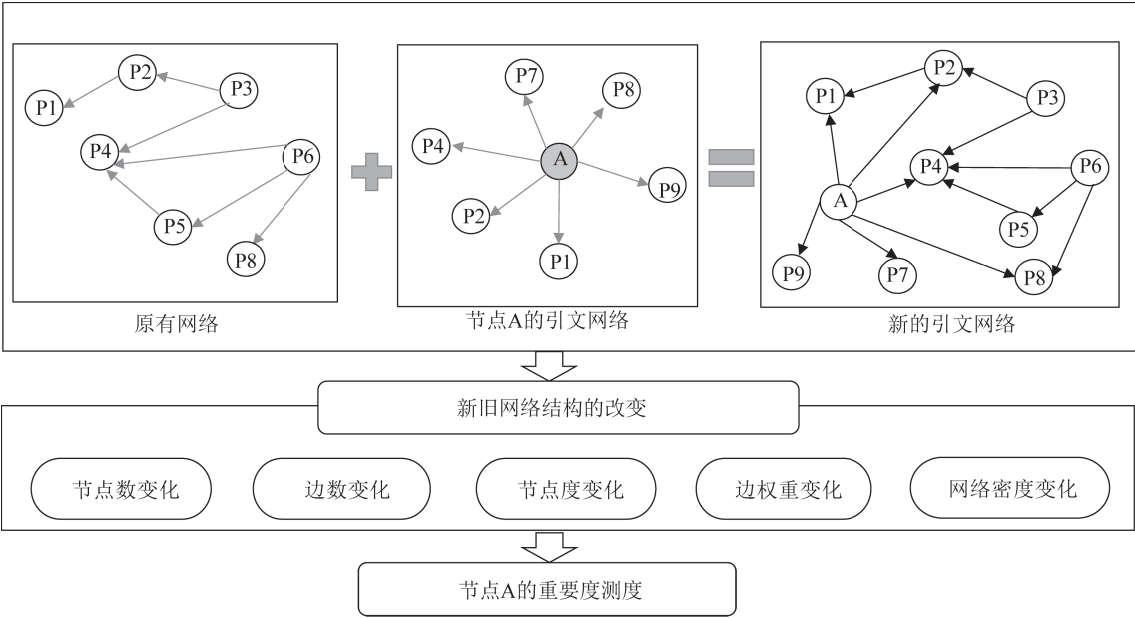


图1 网络节点重要度测度思路方法

网络中新增节点是指原有社区网络中不存在而新加入节点网络中存在的节点。网络中新增边是指原有社区网络节点间不存在的,因新节点网络的加入而重新构建的边,具体包括以下3种情况:①两个节点均不在原网络中存在,两个新加入节点产生边;②原有社区网络中已存在的某个节点与新加入节点网络的某个节点产生边,即仅有一个节点在原网络中存在,另外一个在新节点网络中存在;③两个节点均在原有网络中存在,但不存在关联边,因在新加入节点网络中存在关联而产生边。网络密度通常用来衡量网络中各节点之间的凝聚度,其定义是网络中实际存在的边占所有可能出现边的百分比。网络结构变化指标计算方法见表1。

科研论文所蕴含知识的流动和传承主要通过学术社区内论文之间的相互引证来实现,论文的作者群体是学术社区网络内的引用主体,引用句所蕴含的概念或术

表1 网络结构变化测度相关指标

测量维度	指标名称	计算方法
节点	新增节点数 (Vertex _{add})	叠加融合后的新网络与原有网络节点数量的差异
	新增节点度 (Degree _{add})	叠加融合后的新网络与原有网络节点度的差异
边	新增边数 (Edge _{add})	叠加融合后的新网络与原有网络边数量的差异
	新增边权重 (Weight _{add})	叠加融合后的新网络与原有网络边权重的差异
网络结构	网络密度变化 (Density _{change})	叠加融合后的新网络密度与原有网络密度的差异

语是学术社区内流动的具体知识,重点选取学术社区的作者引用网络和主题引用网络来衡量节点论文的重要度。作者引用网络是基于论文引用关系所构建的论文

作者与参考文献作者之间的网络, 边的方向代表从论文作者到参考文献作者。主题引用网络在对引文文本解析基础上, 通过抽取和标注论文以及所引用文献的重要概念或术语来表示研究主题, 基于引用关系构建引用主题与论文主题之间的网络。

指标具体算法如下:

Input:

$G_{old}(V_{old}, E_{old}) // G_{old}(V_{old}, E_{old})$ 表示基础网络, V_{old} 为基础网络的节点, E_{old} 为基础网络的边;

$G_{node}(V_{node}, E_{node}) // G_{node}(V_{node}, E_{node})$ 表示待测节点网络, V_{node} 为待测网络的节点, E_{node} 为待测网络的边;

Do in Parallel:

$Degree = \text{Sum}(V_i, Degree_i) //$ 计算网络节点的度, i 个节点度之和;

$Edge = \text{Sum}(E_j, Weight_j) //$ 计算网络边的权重, j 条边权重之和;

$Density = m/n(n-1) //$ 计算网络密度, m 为实际存在边数, n 为节点数;

$V_{new} = V_{old} \cup V_{node} //$ 获取基础网络和待测节点网络节点的并集;

$E_{new} = E_{old} \cup E_{node} //$ 获取基础网络和待测节点网络边的并集;

Merged $G_{New}(V_{new}, E_{new}) //$ 合并生成新的网络;

$Vertex_{add} = \text{Minus}(\text{Num}V_{new}, \text{Num}V_{old}) //$ 新增节点数;

$Degree_{add} = \text{Minus}(Degree_{new}, Degree_{old}) //$ 新增节点度;

$Edge_{add} = \text{Minus}(\text{Num}E_{new}, \text{Num}E_{old}) //$ 新增边数;

$Weight_{add} = \text{Minus}(Weight_{new}, Weight_{old}) //$ 新增边权重;

$Density_{change} = \text{Minus}(Density_{new}, Density_{old}) //$ 网络密度变化;

Output:

$Vertex_{add}, Degree_{add}, Edge_{add}, Weight_{add}, Density_{change}$

3 实验数据与过程

3.1 实验数据集构建

近年来新发传染病频发, 新发、突发传染病的防控是国家亟需攻克科技前沿领域之一, 传染病学与流行病学、微生物学、免疫学、寄生虫学和生物化学等多个临床和基础医学领域都具有密切的联系, 相关领域的论文辐射范围较广泛, 对传染病领域重要核心文献的识别有助于快速掌握病毒发展规律, 提升防控能力。因此, 选取传染病领域论文来验证所提出的通过文献网络结构变化测度节点论文重要度方法的科学性和合理性。由于节点论文的重要度是基于引文网络结构变化测度的, 为保证每篇论文有超过3年的引用期, 选用2013—2018年Web of Science数据库中传染病领域(Infectious Diseases)数据为实验数据(数据更新时间为2022年12月20日), 将论文限定在核心合集, 文献类型限定为Article和Review, 共获得62 861篇论文, 并以2013—2017年论文集为基础数据, 构建学术社区原有基础作者引用网络和主题引用网络。以2018年论文集为待测数据, 计算其对2013—2017年原有作者引用网络和主题引用网络结构的改变程度, 进而测度每篇论文的重要度。以2018年论文的引文数据为参照数据, 验证文中方法的合理性。

2013—2017年作者引用网络和主题引用网络的基础网络指标见表2。作者引用网络中节点总数和边总数远大于主题引用网络, 但主题引用网络的平均节点度、平均边权重和总体密度大于作者引用网络。

表2 2013—2017年论文基础网络指标

类别	节点数	边数	总节点度	总边权重	平均节点度	平均边权重	网络密度
作者引用网络	63 001	963 770	1 103 821	232 807	17.52	0.24	0.000 243
主题引用网络	5 554	245 143	218 949	906 469	39.42	0.89	0.007 949

3.2 实验过程

分别构建2013—2018年传染病领域论文集和参考文献集, 在作者名称消歧基础上, 基于引用关系构建2013—2017年作者引用网络和2018年每篇论文的作者引用网络, 分别将2018年每篇论文的作者引用网络与

2013—2017年原有网络进行叠加融合, 对比叠加融合后的新网络与原有网络的结构差异, 分别计算作者引用网络中新增节点数、新增节点度、新增边数、新增边权重和网络密度变化指标值。抽取和标注论文集和参考文献集中标题和摘要所蕴含的概念术语, 去除停用词和高频通用词, 利用所抽取的概念术语来表示每篇论文

的主题内容, 基于引用关系构建研究主题之间的关系网络, 同样将2018年每篇论文的主题引用网络与2013—2017年基础主题引用网络进行叠加融合, 对比新旧主题引用网络结构的差异, 分别计算主题引用网络中新增节

点数、新增节点度、新增边数、新增边权重和网络密度变化指标, 进而基于作者引用网络和主题引用网络结构的变化综合测度每篇节点论文的重要度。具体流程见图2。

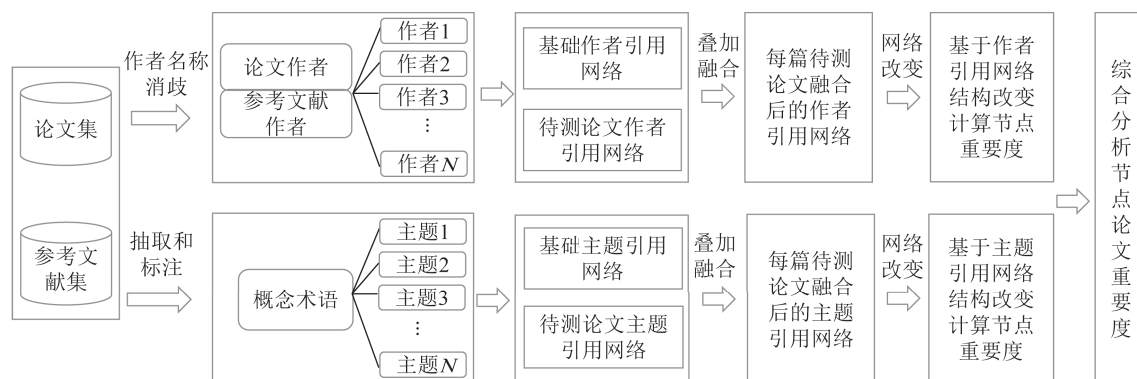


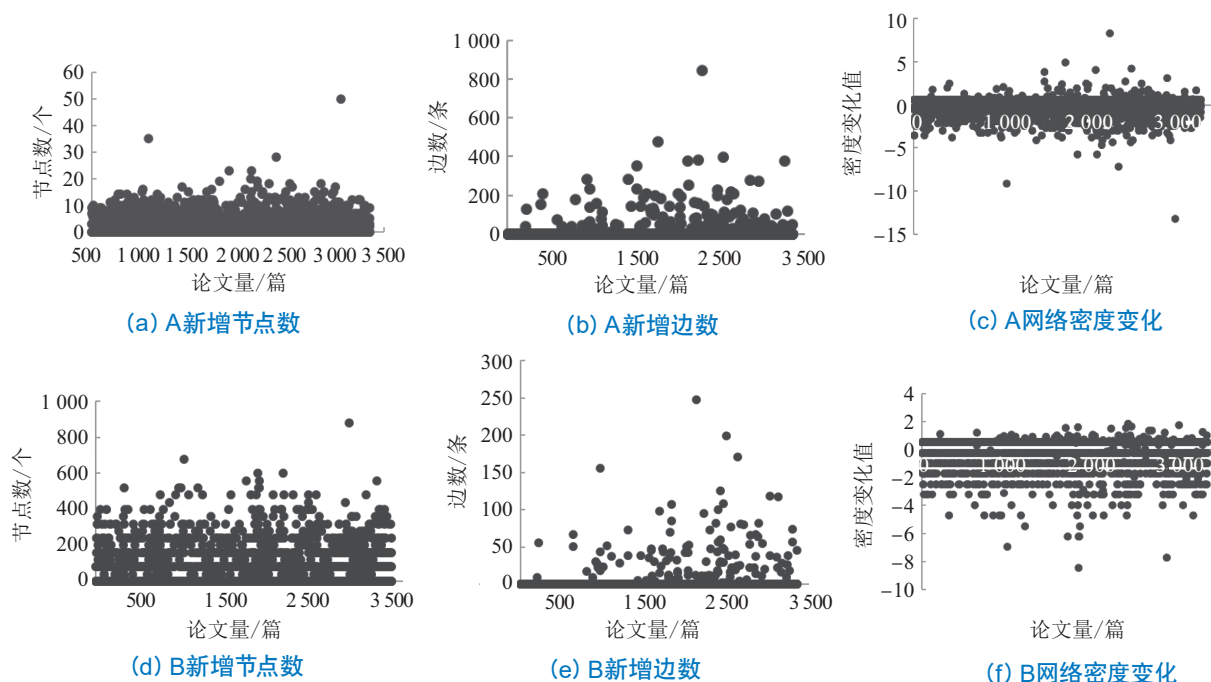
图2 节点论文重要度计算实验过程

4 实验结果分析

4.1 作者引用网络和主题引用网络结构变化的对比分析

节点论文的重要度体现在对原有学术社区内作者引用网络和主题引用网络结构的改变上, 作者引用网络中的节点表示不同学者, 主题引用网络中的节点表示概念术语。由图3可知, 2018年每篇论文自身网络加入后,

2013—2017年基础作者引用网络新增节点数和新增边数的均值分别为2.34和4.57, 而主题引用网络新增节点数和新增边数的均值分别为1.36和1.60, 作者引用网络中新增节点和边的数量均大于主题引用网络, 说明每篇论文网络加入后对基础作者引用网络和主题引用网络结构的改变程度存在差异。虽然随着节点数量的增加, 叠加融合后两个新引用网络的节点和边数量都在增加, 但新增节点并未全部与已有节点建立关系, 网络因新节点的加入反而变得稀疏, 密度降低。



注: A为作者引用网络, B为主题引用网络。

图3 作者引用网络和主题引用网络结构变化指标分析

4.2 节点论文重要度综合得分的计算方法

节点论文的重要度是多个指标的结合,综合考虑节点加入后基础作者引用网络和主题引用网络结构的改变,但每个指标在网络结构特征测度过程中发挥的作用存在差异,应赋予不同的权重。从文中定义的网络结

构特征指标的皮尔逊相关性分析来看,大多数指标之间相关性较强,并具有显著统计学意义,尤其是作者引用网络中新增节点数和新增节点度与新增边权重以及主题引用网络新增节点数和新增节点度之间具有较高的正相关性,新增节点数和新增边数与网络密度变化指标之间为负相关性,统计学意义显著(见表3)。

表3 网络结构变化指标的相关性分析

类别		作者引用网络					主题引用网络				
		新增节点数	新增节点度	新增边数	新增边权重	网络密度变化	新增节点数	新增节点度	新增边数	新增边权重	网络密度变化
作者引用网络	新增节点数	1.00	0.99**	0.67**	0.65**	-0.96**	0.44**	0.44**	0.08**	0.10**	-0.43**
	新增节点度	0.99**	1.00	0.08**	0.08**	-0.96**	0.44**	0.44**	0.09**	0.09**	-0.43**
	新增边数	0.67**	0.08**	1.00	0.97**	0.22**	-0.01	0.01	0.43**	0.38**	0.05**
	新增边权重	0.65**	0.08**	0.97**	1.00	0.22**	0.00	0.00	0.42**	0.38**	0.05**
	网络密度变化	-0.96**	-0.96**	0.22**	0.22**	1.00	-0.43**	-0.43**	0.04*	0.02	0.44**
主题引用网络	新增节点数	0.44**	0.44**	-0.01	0.00	-0.43**	1.00	0.99**	0.10**	0.11**	-0.99**
	新增节点度	0.44**	0.44**	0.02	0.00	-0.43**	0.99**	1.00	0.11**	0.12**	-0.98**
	新增边数	0.08**	0.09**	0.43**	0.42**	0.04*	0.10**	0.12**	1.00	0.98**	0.02
	新增边权重	0.09**	0.09**	0.38**	0.38**	0.02	0.11**	0.12**	0.98**	1.00	0.00
	网络密度变化	-0.43**	-0.43**	0.05**	0.05**	0.44**	-0.99**	-0.98**	0.02	0.00	1.00

注: **表示在0.01级别(双尾),相关性显著; *表示在0.05级别(双尾),相关性显著。

由于指标之间具有较强的相关性,选用主成分因子分析方法,它是基于降维思想在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。主成分因子分析方法用少数的几个变量来综合反映原始变量的主要信息,每个综合指标称为主成分,以每个主成分的方差贡献率为权重,避免了人为赋值带来的主观性。进一步依据每个主成分和权重构造综合评价函数,最终形成综合得分,可以直接按照得分高低来对每篇论文的重要度进行比较和排序。KMO检验和Bartlett球形检验是进行主成分分析前常用的统计学检验方法: KMO检验主要用于验证数据是否适合进行因子分析, Bartlett球形检验则用于检验数据的可分性和相关性。两者检验结果具有统计学意义,说明实验数据适合做主成分因子分析。2018年待测论文集中作者引用网络和主题引用网络结构变化指标的KMO检验值为0.769, Bartlett球形检验结果具有显著统计学意义(见表4),说明适合选用主成分因子分析方法来计算每篇论文加入后作者引用网络和主题引用网络结构改变的综合分值,进而测度每篇论文的重要度。

在利用SPSS软件进行实验数据主成分因子分析过程中,选用相关性矩阵和最大方差方法,提取特征值大

于1的变量作为主成分,共提出3个主成分。第一个主成分特征值为4.288,方差百分比为42.878%;第二个主成分特征值为2.831,方差百分比为28.314%。3个主成分累计方差百分比为87.636%,见表5。每个主成分权重为该主成分的特征值与所选取3个主成分特征值总和的比值。

表4 KMO检验和Bartlett球形检验结果

KMO取样适切性量数		0.769
Bartlett球形检验结果	近似卡方	173 682.625
	自由度	45
	显著性	<0.001

表5 主成分因子分析每个成分的总方差解释

类别	特征值	方差百分比/%	累计方差百分比/%
成分1	4.288	42.878	42.878
成分2	2.831	28.314	71.192
成分3	1.644	16.445	87.636
成分4	0.989	9.891	97.537
成分5	0.224	2.242	99.779
成分6	0.018	0.183	99.962
成分7	0.005	0.052	100.000

主成分载荷矩阵表明了每个原始变量与主成分之间的关联程度,反映了主成分对该原始变量的解释能力。如果某一原始变量在某个主成分上的载荷值较高,则该主成分能够较好地解释该原始变量,反之表明对原始变量的解释能力较低。根据主成分载荷矩阵(见表6)可知,主成分1主要用于解释两个网络中节点和节点度指标的变化,主成分2主要用于解释两个网络中边和边权重指标的变化,主成分3主要用于解释两个网络节点和密度相关指标的变化。

表6 主成分载荷矩阵

类 别		成分1	成分2	成分3
作者引用网络	新增节点数	0.84	0.02	0.53
	新增节点度	0.84	0.02	0.53
	新增边数	0.00	0.85	0.04
	新增边权重	0.00	0.85	0.04
	网络密度变化	-0.82	0.23	0.51
主题引用网络	新增节点数	0.85	0.00	0.52
	新增节点度	0.85	-0.01	0.51
	新增边数	0.12	0.83	0.00
	新增边权重	0.14	0.80	-0.01
	网络密度变化	-0.84	0.09	0.52

按照主成分因子分析方法中综合分值 F 的计算方法,计算每篇论文对学术社区网络的改变程度,通过 F 值来综合判断每篇节点论文的重要度,见式(1)。

$$F = a_1 F_1 + a_2 F_2 + a_3 F_3 \quad (1)$$

式中: F_1 、 F_2 、 F_3 分别为3个主成分的得分; a_1 、 a_2 、 a_3 分别为3个主成分的权重,权重值分别为0.49、0.32、0.19。

4.3 节点论文重要度测度方法的验证

论文被引频次是学术影响力的重要测度指标,为验证基于网络结构特征变化的论文重要度测度方法的科学性和合理性,将由主成分因子分析方法得到的重要度综合得分与2018年待测论文的引文数据进行相关性分析,对重要度综合得分排序与被引频次排序分别进行皮尔逊相关性检验。3种相关性检验方法均证明重要度综合得分与被引频次存在相关性,重要度综合得分排序与被引频次排序也存在相关性,并且检验结果具有显著统计学意义,具体结果见表7。这说明基于网络结构特征变化的重要度测度方法具有可行性。

表7 重要度综合得分和排序与被引频次和排序的相关性分析

重要度综合得分与被引频次		重要度综合得分排序与被引频次排序	
相关系数	显著性(双尾)	相关系数	显著性(双尾)
0.65*	0.005	0.71*	<0.001

注: *表示在0.05级别(双尾),相关性显著。

在2018年实验数据集中,按照重要度综合得分取前1%的61篇论文作为重要论文。实验数据集包含44篇被科睿唯安认定的高被引论文(Highly Cited Papers),其中26篇出现在重要度综合得分前1%论文范围内,说明通过网络结构变化测度论文重要度方法能识别出60%的高被引论文。

5 总结与讨论

科研论文是科学活动产生的主要形式和重要知识载体,含有丰富情报价值,推动了学术研究的发展和传承,不同领域内的高质量、创新性论文更是发挥着承前启后的重要作用,科研论文重要度的评价方法和标准一直是科学界关注重点。科研论文的价值和影响力蕴涵在学术社区网络的复杂交互过程之中,某篇新加入论文对原有学术社区网络结构的改变说明论文对学术社区产生的影响和带来的能量,所产生的新能量和新影响在一定程度上可反映论文的重要度。因此,本文探索通过计算论文对学术社区内作者引用网络和主题引用网络结构的改变程度来测度每篇节点论文的重要度,采用主成分因子分析方法将多个网络指标转换成重要度综合得分。综合得分与公认的被引频次存在较强的相关性,说明提出的方法具有科学性和合理性。

学术社区内作者在交互过程中产生了多种复杂网络,包括合作网络、引用网络、社交网络等。随着科研人员队伍的壮大,新生力量群体不断加入学术社区,学术社区中学者节点数量增多,通过合作或引用关系与原有学术社区内学者形成更为复杂和庞大的网络。很多创新性研究思路和主题需要在原有知识体系基础上经过一定时间的积累和碰撞产生,学者们在短时间内很难快速提出多个新的研究主题,因此主题引用网络中新增节点数小于作者引用网络。未来可进一步分析不同类型网络结构的变化规律,依据各自特征变化设计相应的节点重要度测度指标和方法。不同学科领域科学

发展规律存在差异,目前只选用了传染病领域论文之间的引用网络进行实证,忽略了学科交叉性的影响,此外在主题引用网络中只基于抽取和标注的重要概念或术语来表示引用主题,并未深入引用句进行引文全文文本解析。未来研究将进一步扩大数据实证范围,通过引文全文文本解析构建引用主题网络,进一步验证所提出的节点论文重要度计算方法在不同领域的适用性。同时,引用关系网络是动态网络,同一对象对不同时期网络结构的影响存在差异,基于网络结构变化所得到的重要度也会随时间发生变化,下一步应对比分析不同时期间论文节点的重要度,从而综合考虑论文在整个生命周期中的重要度。

参考文献

- [1] KAUFER D S, GEISLER C. Novelty in academic writing[J]. *Written Communication*, 1989, 6 (3): 286-311.
- [2] 邱均平, 周子番, 张裕晨. 近十年我国计量与评价研究的发展与趋势[J]. *情报理论与实践*, 2021, 44 (2): 192-200, 184.
- [3] TAHAMTAN I, AFSHAR A S, AHAMDZADEH K. Factors affecting number of citations: a comprehensive review of the literature[J]. *Scientometrics*, 2016, 107 (3): 1195-1225.
- [4] BOYACK K W, KLAUVANS R. Predicting the importance of current papers[C]//*Proceeding of the 10th International Conference of the International Society for Scientometrics and Informetrics*. 2005.
- [5] 耿树青, 杨建林. 基于引用情感的论文学术影响力评价方法研究[J]. *情报理论与实践*, 2018, 41 (12): 93-98.
- [6] ERDT M, NAGARAJAN A, SIN S C J, et al. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media[J]. *Scientometrics*, 2016, 109 (2): 1117-1166.
- [7] SATHIANATHAN N J, LANE III R, MURPHY D G, et al. Social media coverage of scientific articles immediately after publication predicts subsequent citations-#SoME_Impact score: observational analysis[J]. *Journal of Medical Internet Research*, 2020, 22 (4): e12288.
- [8] 郭倩影, 赵丹群. 基于Altmetrics的学术论文评价研究评述[J]. *情报理论与实践*, 2021, 44 (9): 190-197.
- [9] 冯婷婷. 学术论文代表作复合式评价方法研究[D]. 重庆: 重庆大学, 2021.
- [10] SHIBAYAMA S, YIN D Y, MATSUMOTO K. Measuring novelty in science with word embedding[J]. *PLoS ONE*, 2021, 16 (7): e0254034.
- [11] WANG Z Y, WANG K Y, LIU J Y, et al. Measuring the innovation of method knowledge elements in scientific literature[J]. *Scientometrics*, 2022, 127 (5): 2803-2827.
- [12] HE J G, CHEN C M. Predictive effects of novelty measured by temporal embeddings on the growth of scientific literature[J]. *Frontiers in Research Metrics and Analytics*, 2018, 3: 9.
- [13] LUO Z R, LU W, HE J G, et al. Combination of research questions and methods: a new measurement of scientific novelty[J]. *Journal of Informetrics*, 2022, 16 (2): 101282.
- [14] CHEN L L, FANG H. An automatic method for extracting innovative ideas based on the Scopus® database[J]. *Knowledge Organization*, 2019, 46 (3): 171-186.
- [15] 曹树金, 闫颂. 基于语义角色信息的科技论文创新段落定位及功能句识别方法研究: 以中文情报学领域论文为例[J]. *情报理论与实践*, 2022, 45 (11): 1-9, 20.
- [16] MISHRA S, TORVIK V I. Quantifying conceptual novelty in the biomedical literature[J]. *D-Lib Magazine*, 2016, 22 (9/10): 9-10.
- [17] 贺婉莹. 基于机器学习的论文学术创新力评价研究[D]. 南京: 南京大学, 2019.
- [18] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic spreading in scale-free networks[J]. *Physical Review Letters*, 2001, 86 (14): 3200-3203.
- [19] KITSACK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. *Nature Physics*, 2010, 6 (11): 888-893.
- [20] ZHANG Y, WANG M, GOTTWALT F, et al. Ranking scientific articles based on bibliometric networks with a weighting scheme[J]. *Journal of Informetrics*, 2019, 13 (2): 616-634.
- [21] MIN C, BU Y, SUN J J, et al. Is scientific novelty reflected in citation patterns?[J]. *Proceedings of the Association for Information Science and Technology*, 2018, 55 (1): 875-876.
- [22] AMPLAYO R K, HWANG S W, SONG M. Evaluating research novelty detection: counterfactual approaches[C]//*Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 2019: 124-133.
- [23] AMPLAYO R K, HONG S, SONG M. Network-based approach to detect novelty of scholarly literature[J]. *Information Sciences*, 2018, 422: 542-557.
- [24] 谭琳洁, 刘向. 面向含时引文网络的论文现时影响力评价[J]. *中国科技期刊研究*, 2020, 31 (4): 468-473.

作者简介

郭红梅, 女, 博士, 副研究馆员, 研究方向: 科学计量、知识组织、文本挖掘。

陈星宇, 女, 硕士, 研究实习员, 研究方向: 科学计量。

李超, 男, 学士, 工程师, 研究方向: 软件工程。

马峥, 男, 博士, 研究员, 通信作者, 研究方向: 情报学、科学计量学, E-mail: mazheng@istic.ac.cn。

Importance Measurement of Scientific Article Based on Effect to the Academic Community Networks Structure

GUO HongMei CHEN XingYu LI Chao MA Zheng

(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract: The inheritance and interaction of knowledge are contained in the interaction process of scholars, which form a complex dynamic network. From the perspective of network evolution dynamics, this paper explores the importance of a node by changing the original network structure, and applies this idea to the literature network to measure the importance of node papers. After overlaying and integrating each test paper network with the original basic network of the academic community, multiple indicators such as node number, node degree, edge number, edge weight, and network density are used to measure the changes in network structure before and after adding each paper. Then the comprehensive score of the network structure indicators is used to measure the importance of node papers. The paper conducts an empirical study based on scientific articles of infectious diseases field from 2013 to 2018 in Web of Science and uses the author citation network and topic citation network from 2013 to 2017 as the basic network of academic communities. Based on basic network, the comprehensive score of the network after adding the citation network for each paper in 2018 is calculated as the importance of the node paper. By referring to the citation data from 2018, it is proven that the importance score is strongly correlated with the frequency of citations.

Keywords: Importance; Research Community; Author Citation Network; Research Topic Citation Network

(责任编辑: 王玮)