



开放科学
(资源服务)
标识码
(OSID)

学科交叉视角下新兴主题识别特征分析 ——以医学信息学为例

杨金庆^{1,2,3} 张力¹

1. 武汉大学信息管理学院 武汉 430072;
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038;
3. 武汉大学图书情报国家级实验教学示范中心 武汉 430072

摘要: [目的/意义] 本文以医学信息学为例探析了学科交叉视角下新兴主题识别特征,对于进一步的新兴交叉学科主题分析具有一定的意义。[方法/过程] 本文首先采用统计分析方法对作者数量、发文量以及资助基金项目进行了历时分析,进一步验证学科交叉主题常规计量维度的适用性;然后运用复杂网络方法分析学科交叉视角下新兴主题的交叉融合新特征以及潜在的网络动态特征。[结果/结论] 实验结果表明:常规性特征发文量、作者数量、资助基金项目数呈指数增长;新兴前阶段学科交叉融合速度较快,篇均医学主题词占比呈显著下降趋势而信息科学呈现快速上升趋势;在新兴前阶段,平均度数 (average degree)、聚集系数 (convergence)、传递性 (transitivity) 指标值均呈现增长趋势;同时,三方关系组分析中 3 边组合的子图数逐年增加,不同学科主题词对的紧密程度明显增强。

关键词: 学科交叉; 新兴主题; PubMed; 医学信息学

中图分类号: G35

The Characteristics Analysis of Emerging Topic Identification from Interdisciplinary Perspective —— A Case Study of Medical Informatics

YANG Jinqing^{1,2,3} ZHANG Li¹

基金项目 富媒体数字出版内容组织与知识服务重点实验室开放基金项目 “多模异构数据环境下新兴主题识别及趋势分析” (ZD2020/09-02)。

作者简介 杨金庆 (1991-), 博士生, 研究方向为数据智能, 信息组织与检索, E-mail: jinq_yang@163.com; 张力 (1992-), 博士生, 研究方向为数据智能, 信息检索与推荐。

引用格式 杨金庆, 张力. 学科交叉视角下新兴主题识别特征分析——以医学信息学为例 [J]. 情报工程, 2021, 7(4): 3-12.

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China;
3. National Demonstration Center for Experimental Library and Information Science Education, Wuhan University, Wuhan 430072, China

Abstract: [Objective/ Significance] This paper explores the characteristics of emerging topics identification from the interdisciplinary perspective with medical informatics as an example, which has some significance for further emerging interdisciplinary topic analysis. [Methods/Process] This paper first employs statistical analysis to analyze the number of authors, publications, and funding programs over time to further verify the applicability of the conventional measurement dimensions of cross-disciplinary topics; and then utilizes a complex network approach to analyze the new features of cross-fertilization and potential network dynamics of emerging topics from a cross-disciplinary perspective. [Results /Conclusions] The experimental results show that: the number of articles, authors, and funding projects of regular features are increasing exponentially; the cross-fertilization of disciplines is faster in the pre-emergence stage, and the percentage of medical topic terms in the average article is decreasing significantly while information science is increasing rapidly; in the pre-emergence stage, the average degree, convergence coefficient, and transitivity are increasing. At the same time, the number of subgraphs with 3-sided combinations in the tripartite relational group analysis increased year by year, and the closeness of subject word pairs in different disciplines increased significantly.

Keywords: Interdisciplinary; emerging topic; PubMed; medical informatics

引言

学科交叉是创新之源。多学科交叉融合一直以来对科学创新起到积极作用。实质上,学科交叉是学术思想的交融、交叉思维方式的综合、系统辩证思维的体现^[1]。早在100多年前,恩格斯就意识到学科之间存在交叉融合的空间,并在《自然辩证法》论著中写道“数学的应用在生物科学中等于零”。随着科学的发展这一想法不断被事实验证,数学向生物学渗透形成了生物数学,物理学原理和方法向生物学渗透创建了分子生物学,还有生物化学、生物力学等交叉学科。由此可见,学科交叉不断推动传统学科的发展,并孕育出新的学科生长点,是当代科学发展的主要趋势之一。

学科交叉是科学发展的内在驱动力。不同

学科间的相互交叉、渗透、融合不断催生新的思想,产生重大原创性研究成果^[2],例如DNA双螺旋结构的发现阐明了生物遗传物质的结构,开辟了分子生物学的新学科领域^[3]。从问题驱动角度,学科交叉是不同的学科面对相类似的科学问题进行研究合作,相互提供研究优势,提升研究方法,寻求解决问题的途径,并建立研究领域新的增长点^[4]。由此可知,将一个学科领域长期发展并已成熟的知识、技术和方法有机地应用到另一学科的前沿领域,在其熔融和撞击的变革中产生重大的创新成果。多学科交叉会聚与多技术跨界融合不断催生新学科前沿、新科技领域和新创新形态^[5]。学科交叉融合不同学科多样化的知识,重塑相应领域的理论和方法,从而产生了新的学科主题。一个学科的形成过程经历了从初期的研究主题、中期

的研究领域,最后才形成具有特定知识体系的新兴学科^[6]。在早期阶段,学科交叉融合是新兴主题形成的内在驱动力之一。深入分析研究主题早期阶段特征有利于更深层次理解新兴主题的形成机制,进而提前引导与国家重大需求相关的研究。

以往新兴主题识别研究从多方面探索了新兴程度计量指标,主要包括发文量^[7-10]、被引量^[11]、作者数量^[12]、基金支持情况^[13-14]常用于计量研究主题的成熟度和被关注程度。这种对多维属性加权融合计算新兴程度的思路被称为“指标与趋势分析法”^[12,15]。常规计量指标仅仅是从科技文献外部特征出发,未就其语义层面的潜在特征进行揭示。库恩^[16]在科学哲学研究中提出新研究主题的出现是建立在已有主题相互作用的基础上,而学科交叉正是多学科相互作用的结果。因此,本研究将立足于学科交叉视角,从宏观外部特征和微观内部特征角度探析学科交叉融合过程中新兴主题的演化特征,为新兴交叉学科主题识别提供可靠的计量维度。

1 研究方案

1.1 数据来源

本文选取具有显著学科交叉特性的医学信息学(Medical informatics)作为研究对象。首先,笔者以检索式 $SU = ('Medical\ informatics')$ 从 web of science 数据库中获取医学信息学研究方向 58601 条记录,数据时间截止于 2019 年 12 月 31 日。其次,考虑到 MeSH 词表是由美国国立图书馆专家委员会维护的、实时更新的叙词表,其收录的主题词对学科发展具有一定的导向性和敏感性,且具有丰富的元数据语义特征和标准的主题词有助于提升研究结果的准确度。将获取的文献映射匹配到 PubMed 数据库,配对成功 54634 条记录,占总记录的 93.2%,数据损失较小。最后,解析美国国立图书馆发布的 MeSH 主题词表,共获得 27336 个主题词(Topic descriptors)作为实验过程中的研究主题,实验涉及数据库表如图 1 所示。

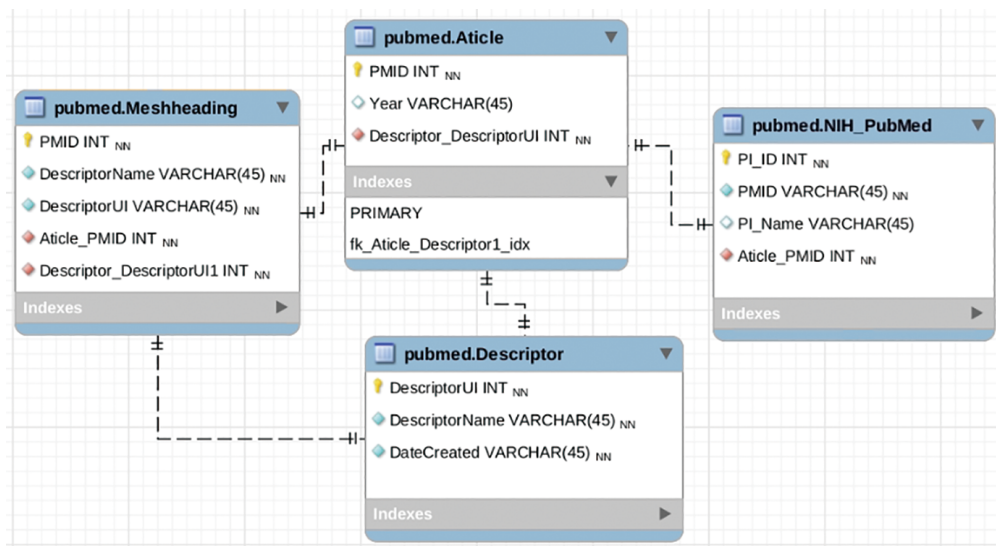


图 1 实验数据结构范例

图1中表Aticle包含文献一般元数据信息如文献唯一编码,标题,出版时间,摘要等。表MeSHheading包含文献唯一编码,Mesh主题词编码及名称。表NIH_PubMed包含文献唯一编码、项目编码、项目名称等。

1.2 研究方法

本文从科技文献外部特征出发,结合学科交叉渗透与网络演化的微观动力特征,借助统计拟合及复杂网络方法分析学科交叉视角下新兴主题的常规特征和交叉融合特征,为尽早识别新兴的交叉研究提供基础,如图2所示。学科交叉视角下新兴主题识别特征的探索既要建立在常规性研究主题新兴程度计量的基础上,又要发掘学科交叉融合过程中的微观动态特征。因此,本文的研究方法立足于常规性指标计量方法,开拓性地验证学科交叉视角下微观动态计量方法的适用性。

(1) 常规性计量方法

科技文献外部特征是识别新兴主题最直观的宏观特征属性,一般包括发布者、发布时间、标

题、关键词、资助金额、被引频次等格式化信息。发文量、被引量、作者数量、基金支持情况常用于计量研究主题的成熟度和被关注程度。根据实验数据特点,本文选取了具有代表性的发文量、作者数量、资助基金项目数来探究其是否适应学科交叉视角下新兴主题的识别。具体而言,本文从宏观层面对选取的常规性指标进行历时统计拟合分析,发掘交叉学科主题的常规属性特征。

(2) 微观动态计量方法

学科交叉的内在驱动是不同学科知识间相互渗透与融合^[17],知识融合是衡量学科交叉程度的重要依据^[18]。再者,学术数据库各种粒度实体间一般存在引文关系、合作关系、共现关系,相应的分析方法为共被引分析、引文耦合分析、作者合作网络分析、关键词共现分析等。这为刻画学科知识交叉融合,相互渗透的动态过程提供了抽象依据。考虑到引文关系的形成需要一定的时间累积^[19],笔者选择即时性较强的共现关系,以MeSH主题词为节点构建主题共现网络,并以网络为媒介探究研究主题交叉融合程度及其共现网络微观动态特征。

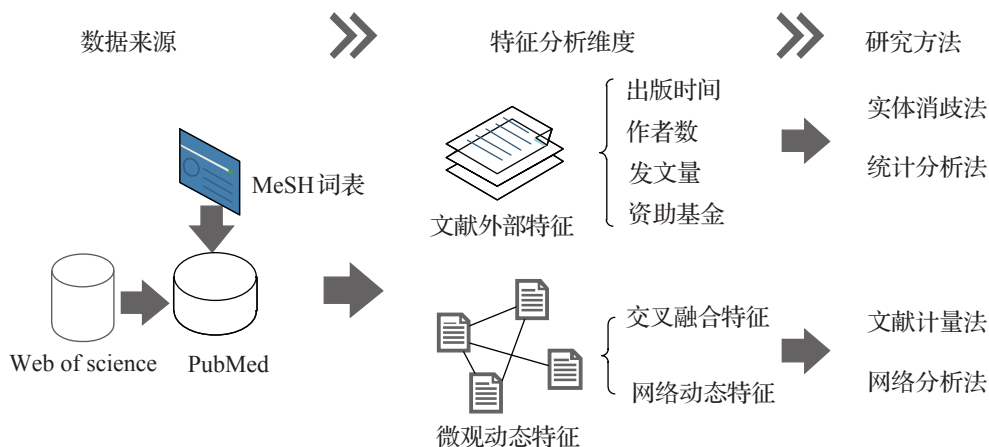


图2 研究方案

总而言之,本文在文献外部特征维度运用统计分析方法对作者数量、发文量以及资助基金项目进行了历时分析,在微观动态特征维度设计了医学类与信息科技类主题词交叉融合紧密度计算方式即两类主题词成对共现频次。网络动态特征分析主要依据复杂网络计算指标来分析网络结构的动态变化。

2 实证及结果分析

依据研究方案从宏观的发文量、作者新增数量、资助基金数等外部特征和学科交叉渗透与网络演化等微观动力学特征角度开展学科交叉视角下的新兴主题特征分析。

2.1 科技文献外部特征分析

依据医学信息学数据集情况,本次实验选取连续发文量超过5篇的年度作为起始年,数据时间跨度为1973-2019年,发文量与作者数量历时演化分析分别如图3、图4所示。

此外,本次实验借助PubMed数据库收录文献的支持基金项目作为学科主题识别的重要特征来源之一。考虑到基金种类多样,为了保证基金项目数据质量,本实验通过抽取文献中资助项目说明数据,筛选医学领域具有代表性和权威性的NIH(National Institutes of Health)资助基金项目作为统计对象,历时统计分析医学信息学获取的NIH基金资助情况,如图5所示。

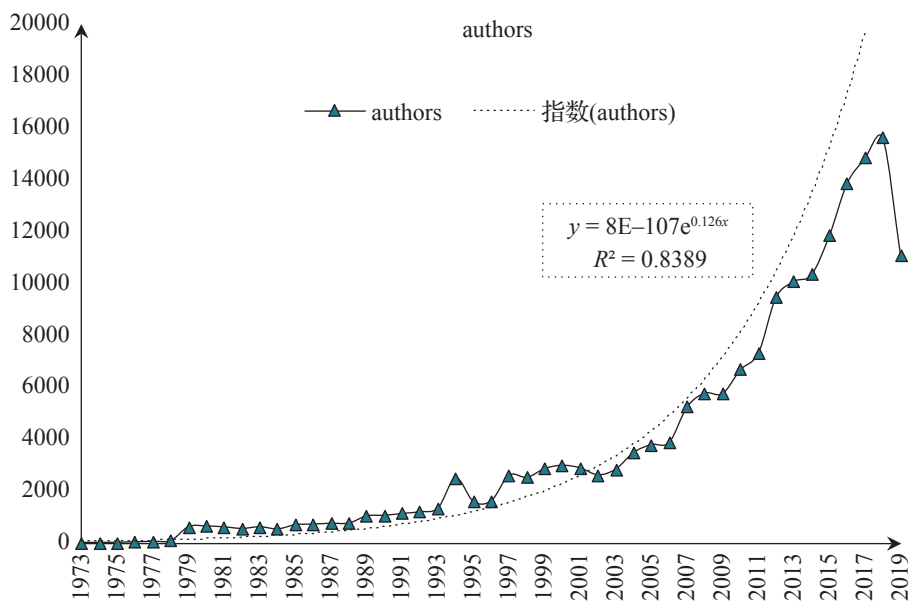


图3 发文量趋势分析

新兴概念一直以来都未能形成统一定义,学科交叉融合达到何种程度便可认为新兴主题也难以确定。本文认为新兴与否受特定领域影响,借助反映其成长情况的各项指标进行量化计算是广泛接受的研究范式,打破研究主题现

有状态的突增现象是判断新兴最为敏感的强信号和直观依据,发掘新兴主题的潜在特征是提升识别效果的弱信号。本部分宏观特征统计是分析学科主题演化趋势的强信号。

医学信息这一主题词收录时间是1987年。

观察发文量、作者数量以及资助基金项目趋势分析图可以发现,1987年之后医学信息这一主题进入快速指数发展的轨道,在1987年之前的一段时间内发文量和作者数量已经在1979年出

现了一次突增现象,且1985年已涉及43项资助基金项目。由此可知,发文量、作者数量以及资助基金项目等外部特征可以作为常规性新兴量化指标来识别学科交叉视角下的新兴主题。

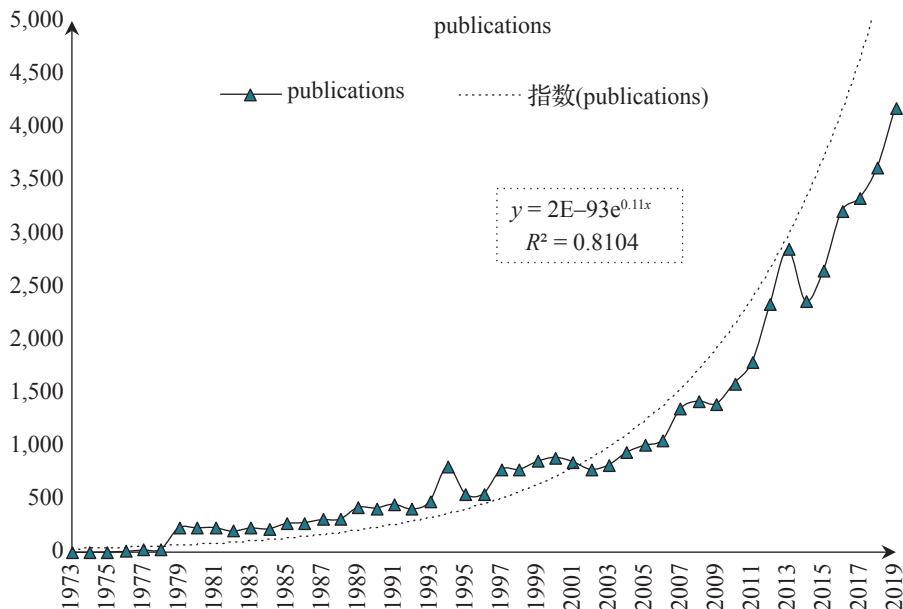


图4 作者数量趋势分析

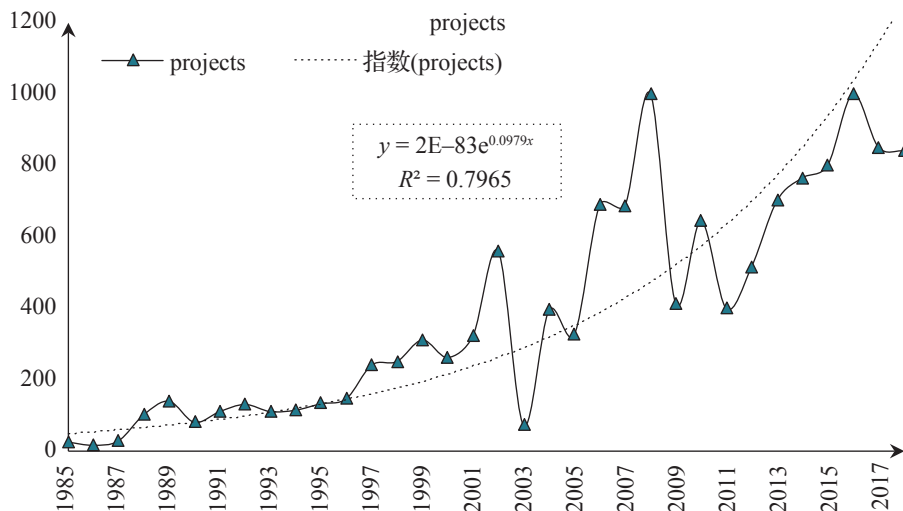


图5 NIH 资助基金项目趋势分析

2.2 学科交叉融合演化分析

医学信息学是由医学和信息科学交叉融合形成的,具有显著的学科交叉特性^[18]。鉴于数据来源于PubMed数据库,笔者通过统计分

析文献中医学类主题词和科技信息类主题词的分布与联系来反映学科主题交叉融合情况。PubMed数据库中的文献包含的主题词主要选取于MeSH词表,词表主要分为A-Z 16个大类,

其中疾病(Disease)、化学物质与药物(Chemicals and Drugs)概念医学特征较为专指,代表性强。

本实验选取了疾病、化学物质与药物两类医学主题词和L类的信息科学(Information Science)作为分析对象。笔者首先统计分析篇

均标注主题词数,然后计算医学类和信息科学类的占比,量化学科交叉融合的过程。

(1) 选取1978-2019年医学信息学相关文献,按年统计文献数量和标引主题词数,进而计算篇均主题词数如图6所示。

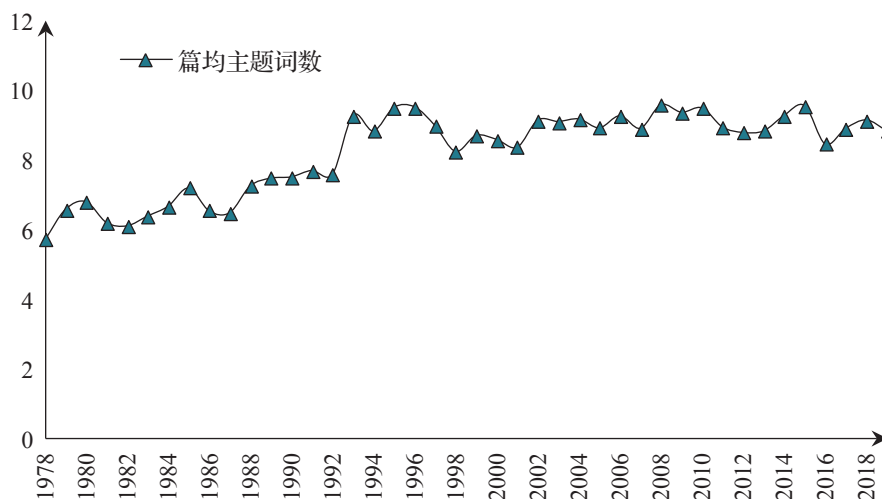


图6 文献篇均主题词数

(2) 由图6可知,文献中的主题词数一般包含8个左右,主题词数相对较为稳定。每篇文献包含主题数量相对恒定的情况下,

计算医学类主题词与信息技术类主题词将能反映学科交叉融合程度,篇均数量演化如图7所示。

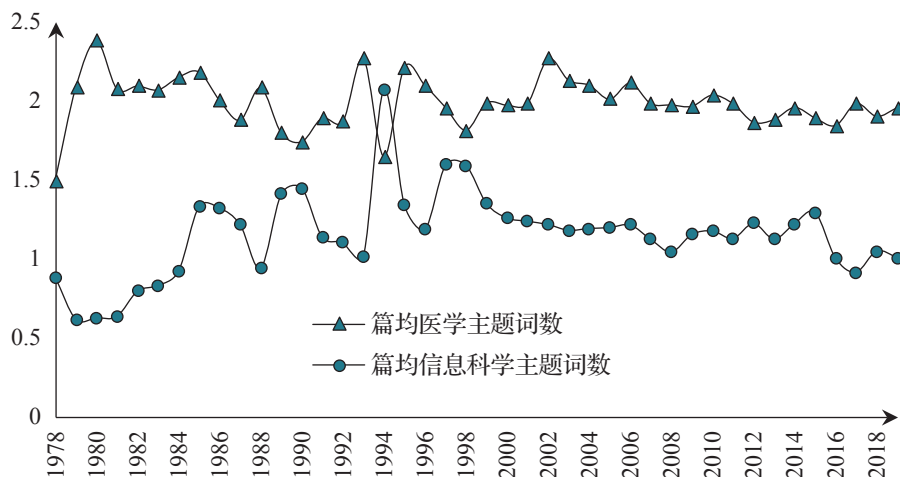


图7 医学类与信息技术类主题词篇均数量

由图7可知,在1987年医学信息学新兴前,学科交叉融合速度较为明显,医学主题词呈显

著下降趋势而信息科学呈现快速上升趋势。学科交叉融合程度可以考虑作为识别新兴交叉主

题的特征之一。

(3) 由于每篇文献主题词数量有限, 计算医学类(C&D)与信息技术类(L)主题词篇均数量存在一定的局限性。信息技术引进或者融合于医学领域, 其目的是为解决当前所面临的重大科研问题, 突破单一学科的局限性。

因此, 两类主题词同时出现在同一篇文献中即增强两者之间融合的程度, 随着不断的融合特定技术将会有效解决特定的医疗问题, 进而形成解决此类问题的一种新范式。笔者通过计算C-L&D-L主题词对的紧密程度进一步补充验证(2)中所的结论, 实验结果如图8所示。

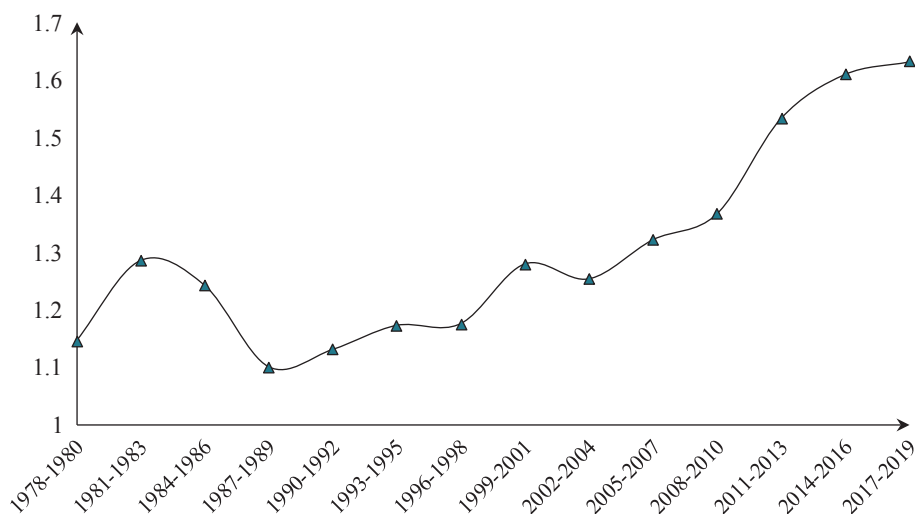


图8 学科交叉紧密程度演化分析

本实验以3年为时间跨度历时分析不同时间段内C-L&D-L主题词对的紧密程度, 以医学类和信息科技类主题词同时出现的文献次数, 然后计算同一时间段均值来衡量学科交叉融合程度。已知医学信息主题词新兴于1987年, 结合图8可知, 在医学信息主题新兴前交叉融合程度呈现明显突增, 这种现象印证了(2)中的实验结论, 同时也说明了交叉学科主题紧密程度可以作为识别新兴主题的潜在特征。

2.3 网络动态特征分析

学科交叉点往往是科学新的生长点、新的科学前沿, 产生重大的科学突破, 使科学发生革命性的变化。学科主题一般包含不同的研

究领域和应用相关的概念, 特定学科主题下的不同概念间相互作用, 相互渗透、融合。本文结合实验数据集特征, 选择通过关键词共现关系构建概念间的相互作用网络, 借助网络统计性质描述网络的拓扑性质。由于学术概念为节点的连边表示相互作用, 因此2.2节统计描述包含了动力学成分。首先, 本实验在构建时序网络的基础上, 计算无向网络的静态特征并进行历时分析, 剖析概念网络相互作用过程中蕴含的动态特征及动力学规律对识别新兴主题的潜在作用。医学类与信息技术类概念主题所构成的动态网络中平均度数、网络密度、聚集系数以及传递性指标值的演化趋势分析如图9所示。

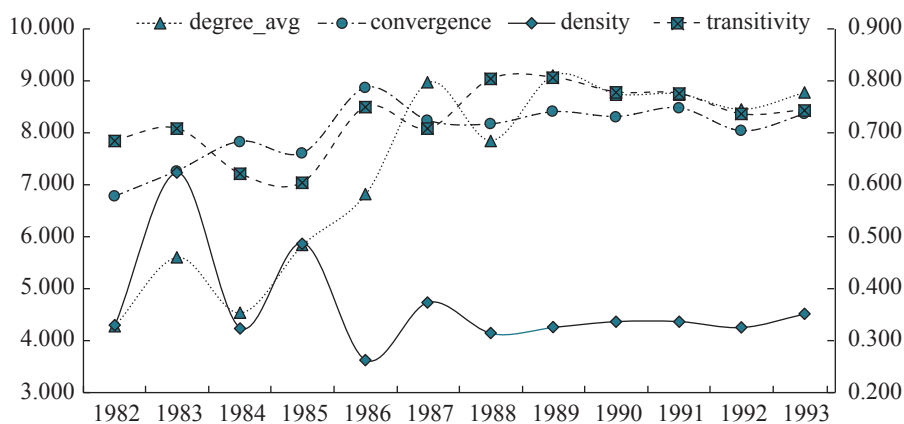


图9 复杂网络特征分析

由图9所示,除网络密度(density)外,平均度数(degree_avg)、聚集系数(convergence)、传递性(transitivity)指标值均在医学信息学新兴时间点1987之前呈现明显增长趋势,尤其网络平均度数,而后出现平稳状态。

新兴前与新兴后增长状态存在明显差异。为了进一步发掘网络动态特征,本文分析了时序网络的三方关系组的动态演化,利用首、尾年增长率($growthindex = H_i^{first} - H_i^{last} / H_i^{first}$)来衡量增长趋势,计算结果如表1所示。

表1 动态网络结构三方关系组历时分析

新兴后阶段					新兴前阶段				
年份	H_0	H_1	H_2	H_3	年份	H_0	H_1	H_2	H_3
1988	1763	495	297	45	1982	224	104	32	4
1989	2420	761	398	75	1983	51	32	31	6
1990	1895	650	315	65	1984	294	105	52	4
1991	1906	628	326	65	1985	141	90	46	9
1992	1883	714	273	55	1986	2118	490	291	26
1993	1634	628	274	64	1987	1402	561	284	53
增长率	-0.073	0.269	-0.077	0.422	增长率	5.259	4.394	7.875	12.250

表1中 H_0 表示孤立节点数, H_i 表示由 i 条连边组成的子图数。观察表中数据可以看出,3边组合的子图数在新兴前阶段逐年增加,而新兴后阶段趋势缓和。计算比较两个阶段的增长率可以发现,新兴前阶段增长趋势明显。医学信息学科主题在新兴前具有较强的增长趋势。

3 总结

学科交叉创造知识是现代科学不可或缺的研究模式。

交叉学科环境下对新兴主题的识别提出了新的挑战。探究学科交叉研究的演化特征可以进一步剖析交叉学科环境下新兴主题特征。

首先,本文宏观统计分析了科技文献外部特征,进一步验证常规性新兴主题量化指标的适应性。统计发现发文量、作者数量、资助基金项目数呈指数增长。其次,通过统计文献中医学类主题词和信息科学主题的情况来反映交叉融合情况,计算医学类与信息技术类主题词篇均数量和C-L&D-L主题词对的共现次数来衡量医学信

息学的交叉融合程度。分析发现医学信息学新兴前, 学科交叉融合速度较为明显, 篇均医学主题词占比呈显著下降趋势而信息科学呈现快速上升趋势, C-L&D-L 主题词对的紧密程度明显突增。因此, 学科交叉融合程度可以考虑作为识别新兴交叉主题的特征之一。最后, 本文结合实验数据集特征, 选择通过关键词共现关系构建概念间的相互作用网络, 借助网络统计性质描述网络的拓扑性质。平均度数 (degree_avg)、聚集系数 (convergence)、传递性 (transitivity) 指标值均在医学信息学新兴前呈现明显增长趋势。同时, 三方关系组分析结果显示 3 边组合的子图数在新兴前阶段逐年增加, 医学类与信息科学类概念间的作用紧密度增强。

由多维分析可知, 常规性科技文献外部特征仍适用于新兴交叉学科主题的识别, 学科交叉融合是已有研究主题相互作用的结果, 量化概念间的交叉作用可用于新兴主题识别指标的制定。本文从网络动态层面探究新兴前后网络状态的差异性, 进而启发未来新兴主题识别研究。

参 考 文 献

- [1] 路甬祥. 学科交叉与交叉科学的意义 [J]. 中国科学院院刊, 2005(1):58-60.
- [2] 张军. 对促进学科交叉, 推动创新发展的认识和思考 [J]. 中国基础科学, 2013, 15(6):42-46.
- [3] 周光召. 发展学科交叉 促进原始创新——纪念 DNA 双螺旋结构发现 50 周年 [J]. 物理, 2003(11):707-711.
- [4] 刘正中. 学科交叉下的术语创新 [J]. 人文论丛, 2009(1):50-64+5.
- [5] 吴朝晖. 以学科交叉融合服务国家战略需求 [N]. 人民日报, 2020-11-04(12).
- [6] Braun T, Schubert A P, Kostoff R N. Growth and trends of fullerene research as reflected in its journal literature[J]. Chemical Reviews, 2000, 100(1): 23-27.
- [7] Tu Y N, Seng J L. Indices of novelty for emerging topic detection[J]. Information Processing & Management, 2012, 48(2): 303-325.
- [8] 殷蜀梅. 判断新兴研究趋势的技术框架研究 [J]. 图书情报知识, 2008(3):76-80.
- [9] 范云满, 马建霞. 基于 LDA 与新兴主题特征分析的新兴主题识别研究 [J]. 情报学报, 2014, 33(7):698-711.
- [10] 杨海霞, 高宝俊, 孙含林. 基于 LDA 挖掘计算机科学文献的研究主题 [J]. 现代图书情报技术, 2016(11):20-26.
- [11] Wang, Q. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology, 2018, 69(2):290-304.
- [12] Rotolo D, Hicks D, R. Martin B. What is an emerging technology? [J]. Research Policy, 2015, 44(10):1827-1843.
- [13] 徐路路, 王效岳, 白如江, 等. 基于 DTM 模型和文本特征分析的基金项目新兴趋势识别研究——以 NSF 石墨烯领域为例 [J]. 数据分析与知识发现, 2018, 2(3):87-97.
- [14] 王效岳, 刘自强, 白如江, 等. 基于基金项目数据的研究前沿主题识别方法 [J]. 图书情报工作, 2017, 61 (13):87-98.
- [15] Porter A L, Garner J, Carley S F, et al. Emergence scoring to identify frontier R&D topics and key players[J]. Technological Forecasting and Social Change, 2019(146):628-643.
- [16] Kuhn, T S and Hawkins D. The structure of scientific revolutions. American Journal of Physics, 1963, 31(7):554-555.
- [17] 刘晓娟, 李广建, 化柏林. 知识融合: 概念辨析与界说 [J]. 图书情报工作, 2016, 60(13): 13-19.
- [18] 齐燕, 许海云, 方曙. 基于 WOS 数据的医学信息学学科交叉发展态势研究 [J]. 中华医学图书情报杂志, 2016, 25(11):30-41.
- [19] Mariani M S, Medo M, Zhang Y C. Identification of milestone papers through time-balanced network centrality[J]. Journal of Informetrics, 2016, 10(4):1207-1223.