



开放科学
(资源服务)
标识码
(OSID)

近五年信息素养主题演化及影响力分析

吴帼帼

山东大学图书馆 济南 250100

摘要: [目的/意义] 信息素养与生活学习工作密切相关, 探究信息素养的发展脉络和研究主题分布以及演化关系, 对促进信息素养培育发展以及指导信息素养能力提升具有积极作用。[方法/过程] 基于近五年 SCI、SSCI 中的信息素养相关研究数据集合, 使用 LDA 模型挖掘各个年代下相关研究的主题分布以及主题间的演化规律, 与此同时基于替代计量学指标找到近五年高社会影响力的具体文章。[局限] 未将信息素养相关的中文期刊文献等数据包含其中。[结果/结论] 近五年信息素养研究主要集中于信息素养、学术图书馆、信息素养教育、数字数据素养、虚假信息以及健康素养等方面。且主题之间相似度较高, 主题衍生关系和延续性较强。

关键词: 信息素养; 相关性分析; LDA; Altmetrics; python

中图分类号: G35

Analysis on the Theme of Information Literacy in Recent Five Years

WU Guoguo

Shandong University, Jinan 250100, China

Abstract: [Purpose/Significance] Information literacy is closely related to life learning, especially in the context of a special era, exploring the development context of information literacy, the distribution of research topics, and the evolutionary relationship plays a positive role in promoting the cultivation and development of information literacy and guiding the improvement of information literacy capabilities. [Method/Process] Based on the information literacy related research data set in SCI and SSCI in the past five years, use the LDA model to mine the topic distribution of related research in each era and the evolution rules between topics. At the same time, find specific articles with high social influence in the past five years based on alternative econometric indicators. [Limitation] Data such as Chinese journals and literatures related to information literacy are not included. [Results/Conclusion] In the past five years, the theme of information literacy has been concerned, mainly focusing on information literacy, academic libraries, information literacy education, digital data literacy, false information and health literacy. The similarity between topics is high, and the derivative relationship and continuity of topics are strong.

Key words: information literacy; Correlation analysis; LDA; Altmetrics; python

基金项目 山东大学研究生教改项目“新文科建设背景下研究生信息素养教育改革研究”(XYJG2020174); 山东大学图书馆项目“新工科创新服务团队”(2022CXTD03); 山东大学图书馆项目“新文科创新服务团队”(2022CXTD02)。

作者简介 吴帼帼(1992—), 硕士, 馆员, 研究方向为知识服务与信息素养, E-mail: guoguow@sdu.edu.cn。

引用格式 吴帼帼. 近五年信息素养主题演化及影响力分析[J]. 情报工程, 2023, 9(1): 92-101.

引言

美国大学与研究图书馆协会 ACRL2015 年出台《高等教育信息素养框架》(framework for information literacy for higher education) 并将信息素养的定义为对信息的反思性思考并利用信息创造新知识的综合能力^[1], 强调信息素养对其他实践行为等影响^[2]。美国图书馆学会 CILIP^[3] (the library and information association) 2018 年将信息素养定义更新为: Information literacy is the ability to think critically and make balanced judgements about any information we find and use. It empowers us as citizens to develop informed views and to engage fully with society。更新后的定义强调了信息素养在五个关键环节中的价值: 公民身份(了解周围世界)、每天(在网上查找信息)、教育(教育阶段培养批判性思维)、工作(有助于提高就业)、健康(寻找可靠的信息来源)。中国图书馆学会 2021 年 9 月引发的《中国图书馆学会“十四五”发展规划纲要(2021-2025 年)》^[4] 中“全民阅读促进于专业指导中心建设”章节中指出, “通过促进全民阅读, 普及图书馆学、情报学和信息技术等基本知识, 提高社会公众的图书馆意识与信息素养”“引导业界创新推进人文素养、艺术素养、科学素养、信息素养、数据素养、职业素养、金融素养、媒体素养等教育”。信息素养从提出到现在一直受到业界的广泛关注, 在该背景下, 挖掘分析信息素养的研究脉络和发展走势, 可以更好的揭示其发展全貌, 对于信息素养相关研究都具有促进意义。因此, 本文首先通过 LDA 模型、python 可视化工具以及

桑基图方法, 析出近五年信息素养相关研究中的主题并可视化分析主题之间的演化情况; 与此同时, 使用 Altmetrics 替代计量学指标^[5] 分析近五年高社会影响力的文献。结合两方面分析, 给出信息素养主题下的高关注度主题, 以及高社会影响力的科研文献, 对信息素养相关研究具有重要的理论和实践意义。

1 数据源的获取

本研究的数据集是基于 Web of Science 核心合集数据库(以下简称为 WoS)构建的, 考虑到跨学科、学科交叉研究, 即非图书情报学科领域也会有信息素养主题相关的研究文献, 所以在实施检索时选择 WoS 核心合集数据库中的 SCI-EXPANDED、SSCI 作为检索数据库, 检索关键词使用 information literacy, 检索式为 TS=(“information literacy”), 限定出版日期为 2017 年 1 月 1 日—2021 年 12 月 31 日, 检索结果为 1 167 条记录。文献类型包含论文、会议录和综述论文等类型。论文年代分布如图 1 所示。

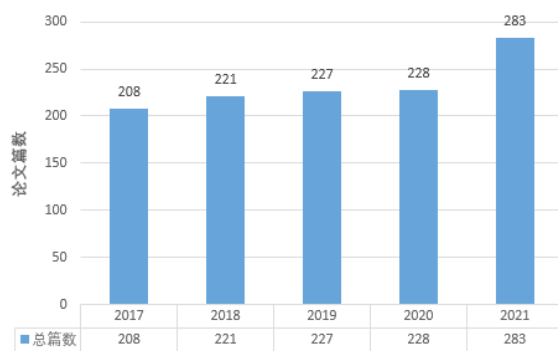


图 1 信息素养主题论文年代分布

2 数据处理方法

主题抽取采用的 LDA 模型。LDA 模型

即隐含狄利克雷分布 (Latent Dirichlet Allocation, 简称 LDA), 由 Blei, David M. 等人 2003 提出的, 可将科研文献数据集中的每篇文章, 以概率分布的形式给出主题分布情况^[6]。LDA 在主题模型中占有重要地位, 可以有效分析大规模文档数据集^[7-9]。需要注意的是, 使用 LDA 模型进行主题分析首先需要预先设定主题个数, 然后在设定的主题个数下根据概率模型得到主题的划分。目前主题个数的确定使用较多的有一致性检验 Coherence、困惑度 Perplexity 分析两种方法。主题一致性主要是衡量主题内聚类效果主题质量。主题困惑度表示在文本分析中, 训练出来的模型识别文档中的主题分布时具有的不确定性。Blei^[10] 就曾在论文中实验中用 Perplexity 困惑度值作为主题个数的评判标准。主题一致性数值越高、困惑度数值越低, 主题内聚类相关性越高^[11]。主题与主题之间的演化分析采用的是余弦距离, 通过两两计算相邻主题之间余弦距离确定主题相似度。余弦距离值的计算是以向量空间中两个主题向量的夹角的余弦值为衡量标准, 余弦距离 = $1 - \text{余弦值}$ 。

余弦距离越接近 1, 表示两个主题的向量夹角的余弦值越小, 即两个向量的夹角越小, 两个主题之间的相似度就越高。

具体数据处理方法为: 首先针对信息素养主题的文献数据集按照年代区间划分; 然后分词预处理文档集合, 设置数据字典处理学科领域的专业词汇, 将 “information literacy、IL、IR” 等词语作为固定词组进行分词, 同时通过使用 python-NLTK 自然语言处理包以及设置 stopWords 停词表, 对常见的英文虚词介词、常见标点符号以及结构化摘要的标注性词语 (如 Objective、Purpose、Design、methodology、approach、Findings、Originality) 做停词处理, 即分词时跳过这些词语, 不参与分词结果统计; 在分词结果数据集的基础上, 通过计算不同主题个数下的困惑度、一致性数据并可视化, 结合两个曲线走势图, 得到较优的主题个数。最后计算相邻主题之间的余弦距离, 得到相邻年代区间不同主题之间的相似度, 进行主题演化可视化分析。具体数据处理流程如图 2 所示。

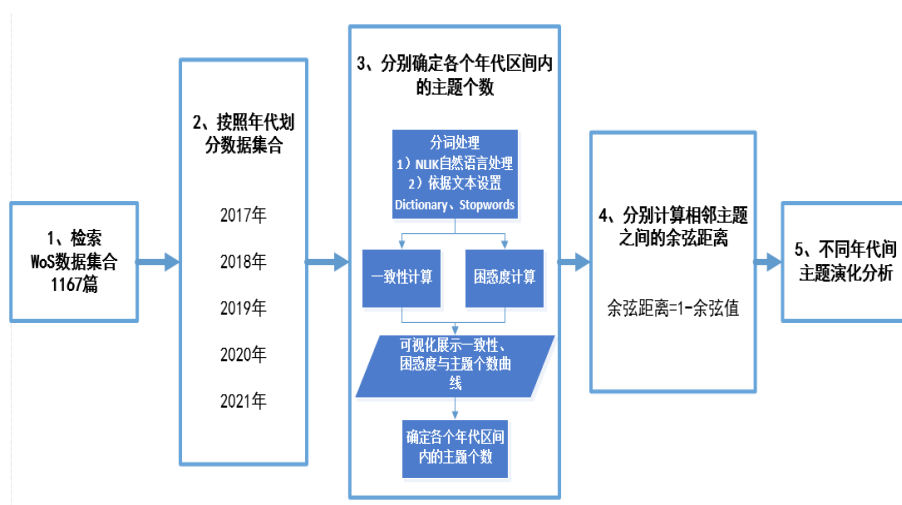


图 2 原始数据处理流程

3 基于 LDA 模型的主题演化分析

为对近五年信息素养相关研究进行主题演化分析,首先通过计算主题一致性、主题困惑度,得到主题个数与主题一致性、困惑度曲线如图3所示,选取主题一致性上升、困惑度下降的交汇点作为确定主题个数的参考点,当曲

线出现两次交汇点的时候,分别通过 pyLDA 可视化展示主题高频词,对比分析确定该年代下的最终主题个数。由各个年代的主题一致性曲线、主题困惑度曲线图可得到,各年代主题个数分别为2017年8个主题、2018年8个主题、2019年6个主题、2020年8个主题、2021年9个主题。

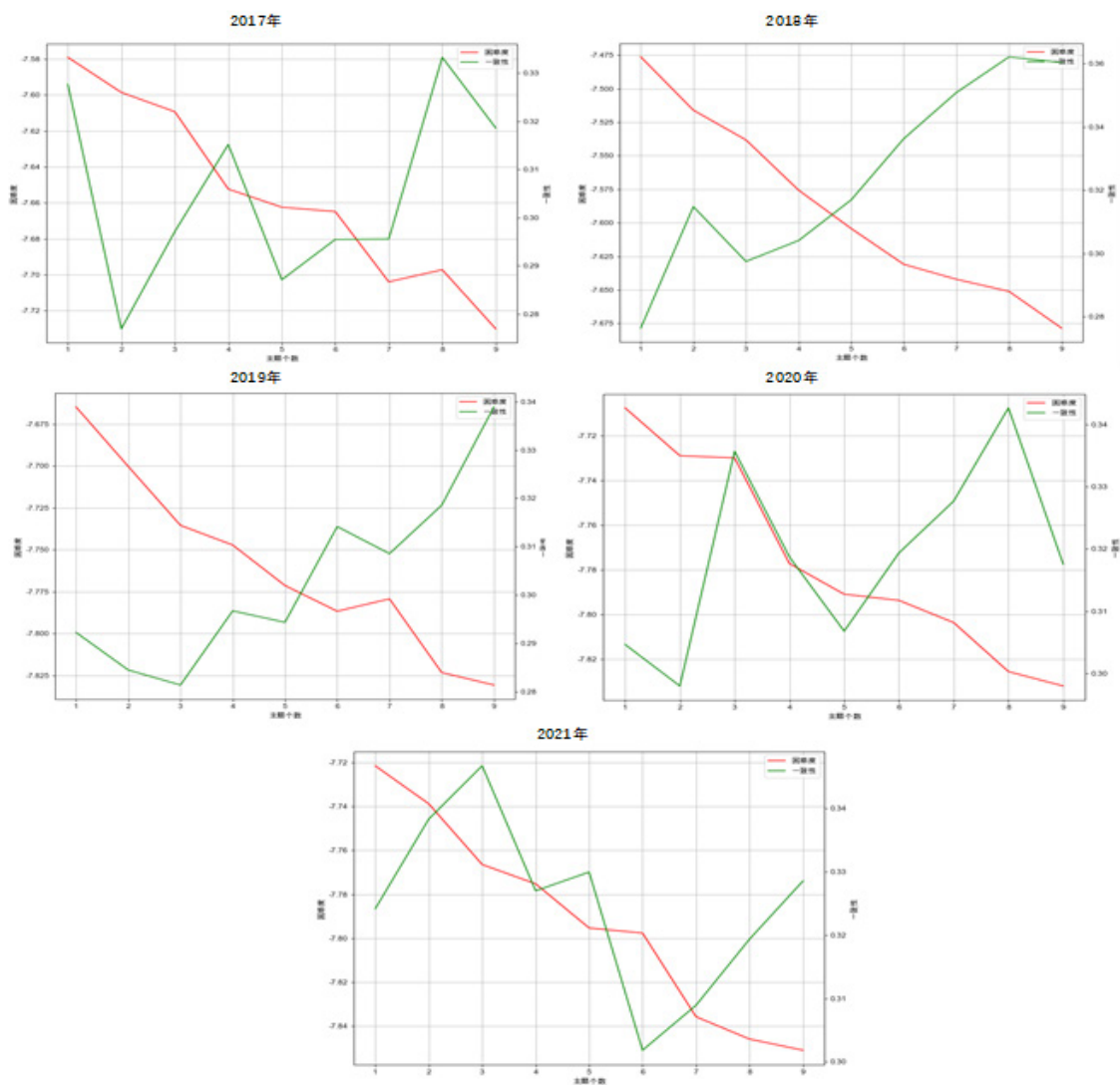


图3 不同年代区间主题困惑度、一致性曲线

结合确定的各个年代主题个数,使用 python 的可视化程序可视化展示各个年代区间下的主题词频分布图如图4所示,结合根据可视化结果分

别查看某主题下的最相关专业术语(top-30 Most Relevant Terms),概况确定各年代各个主题名称。具体的年代一主题名称分布情况如表1所示。

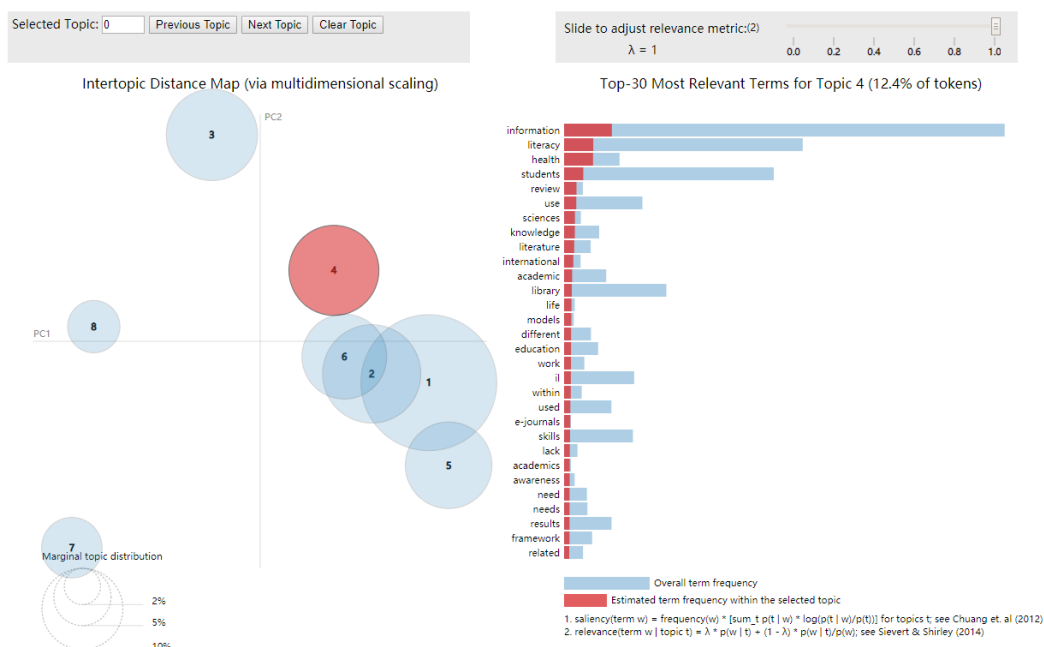


图 4 2017 年信息素养研究主题词频图

表 1 各年代区间内主题高频词分布表

年份-主题	高频词	主题名
2017-topic1	information . literacy	信息素养
2017-topic2	students . information . skill	学生信息技能
2017-topic3	ict . electronic . il	信息电子技术
2017-topic4	health . chealth . knowledge	健康知识
2017-topic5	librarian . college . course	大学图书馆员
2017-topic6	fake . news	虚假信息
2017-topic7	makerspace . faith	创客创业
2017-topic8	behavioral . enjoyment . dgbl	网络社交行为
2018-topic1	information . literacy	信息素养
2018-topic2	librarians library	图书馆员
2018-topic3	information . health . literacy	健康素养
2018-topic4	students . course	课程建设
2018-topic5	digital . skills . literacy	数字素养
2018-topic6	students . education . teaching	教学培训
2018-topic7	online . course	在线课程
2018-topic8	library . academic	学术图书馆
2019-topic1	information . literacy	信息素养
2019-topic2	students . library . librarians	图书馆
2019-topic3	news . theory . internet	网络信息
2019-topic4	social . literacy . academic	社会素养

年份-主题	高频词	主题名
2019-topic5	digital . education	数字教育
2019-topic6	teaching . methods . management	教学方法
2020-topic1	information . health . literacy	健康素养
2020-topic2	information . literacy	信息素养
2020-topic3	poverty . news . fake	信息匮乏
2020-topic4	students . competencies . skills	信息技能
2020-topic5	library . online . education	图书馆在线教育
2020-topic6	librarians . academics	馆员学术能力
2020-topic7	digital . literacy . DL	数字素养
2020-topic8	library . covid-19	疫情
2021-topic1	information . health . literacy	健康素养
2021-topic2	students . online . university	大学生线上教育
2021-topic3	literacy . digital . data	数字数据素养
2021-topic4	health . misinformation	健康虚假信息
2021-topic5	literacy . digital . academic	数字学术素养
2021-topic6	digital . skills . online	数字技能
2021-topic7	digital . health . knowledge . covid-19	健康知识
2021-topic8	librarians . literacy . skills	馆员素养
2021-topic9	information . refugees . sociocultural . barriers	信息难民

根据各个年代区间析出的主题，分别计算相邻年代区间不同主题之间的余弦距离值，根据相邻窗口主题间的余弦距离值，结合相关研究确定 0.3 为相似度阈值^[12]，绘制主题演化的桑基图如图 5 所示。每个元素都对应年代区间

内的某个主题，主题之间的连线代表主题间的流动方向是由左侧主题流向右侧主题。主题与主题之间连线越粗代表两个主题余弦距离值越高，即主题间的相似度越高，两个主题之间的演化关系越紧密。

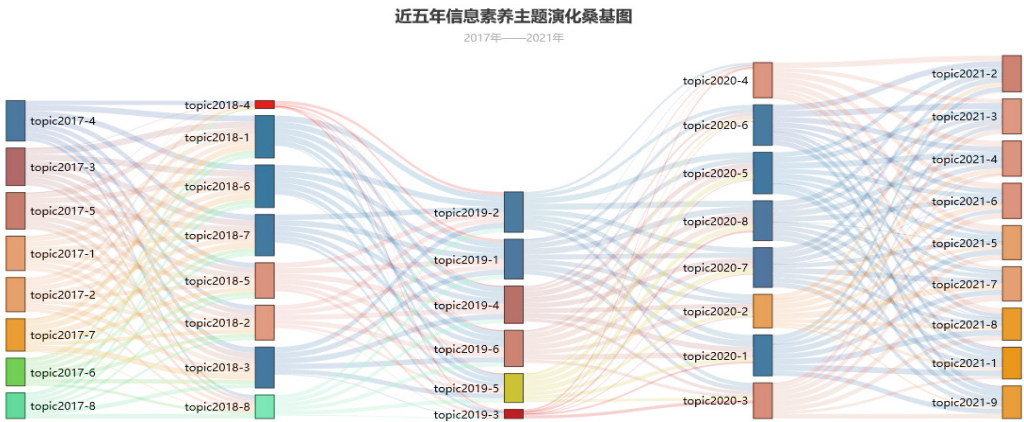


图 5 主题演化桑基图

综上所述,最近五年信息素养在各个学科的研究,主要集中于信息素养、大学图书馆、学术图书馆、馆员素养、信息素养教育、数字数据素养、虚假信息以及健康素养等方面。结合主题演化桑基图可知,信息素养相关研究的主题与主题之间联系较为密切,主要表现为主题间的相似度较高,衍生演化关系和延续性较强。

具体来看:(1)信息素养教育教学一直是该领域的研究重点和热点,表现在各个年代区间下均有如学生信息技能、课程建设、教学培训、在线课程、教学方法、在线教育以及大学生线上教育等相关主题,且主题之间的演化关系较强。尤其是在线课程主题在2018年出现后,2019年未出现,但是在近两年又以图书馆在线教育、大学生线上教育引发关注,即新时期特殊背景下图书馆如何进行在线的信息素养教育教学,大学生如何参与线上教育教学。

(2)数字数据素养方面,数字素养、数据素养在中文意思理解上偏差并不大,但是从主题词分布来看,数字素养为digital literacy、数据素养为data literacy,是两个不同主题。目前关于数字素养、数据素养的概念划分并不是很清晰,部分学者研究认为数字素养教育应包含科学数据素养内容^[13]。按照年代主题演化来看,2018年出现digital skills数字技能主题,2019年digital education数字教育,2020年转化为digital literacy(DL)数字素养主题,2021年digital literacy、data literacy数字素养、数据素养两个主题,digital和data并列出现在主题中,与此同时academic digital literacy数字学术素养,online digital skills数字技能等。从主题之间的

连线粗线程度来看主题间的相似度,可以发现主题之间关联密切,该主题的相关研究延续性较强。(3)健康素养方面,可以看出近五年的研究中占到比较重要的地位。从2017年的health、ehealth、knowledge健康知识主题,演化到2018年的health information literacy健康信息素养,一直发展到近两年的health literacy健康素养,以及2021年出现health misinformation健康虚假信息识别、digital health knowledge、covid-19疫情下健康知识等主题。(4)值得注意的是,虚假信息、社会素养在近五年的相关研究积累中,在2021年衍生出现了新的主题信息难民(information、refugees、sociocultural、barriers),即在原有虚假信息识别、网络社交行为以及社会素养、信息匮乏等研究主题的发展中,大家逐步意识到信息素养、信息获取与利用差异对社会生活的影响,从而上升到新高度,出现信息难民这个主题。

4 近五年信息素养研究的社会影响力分析

科研成果影响力目前主要分为学术影响力和社会影响力两个方面,由于科研评价奖励大多与影响因子、分区以及文献的引用次数相关,所以学术影响力往往受这些因素影响。但仅依靠引文空间分析科研影响力,不可避免的会出现人文因素以及学科差异等问题。在该背景下,加入替代计量学指标进行社会影响力分析可以均衡人为因素产生的引用量过高、学科差异大等问题。以Altmetric、PLoS ALM、impStory、plum analytics等为代表的替代计量学指标逐步被业界关注^[14]。2015年邱均平^[15]明确指出,

Altmetrics 指标与传统计量评价指标可相互补充揭示科研成果的影响力。该指标是评价学术论文在学术界以外的像是社交媒体上造成的影响力,其通过在线社交网络对学术文献的推荐收

藏、谈论和分享等行为体现科研成果的社会影响力^[16]。以 Altmetric 指标分值为衡量社会影响力依据,得到近五年信息素养领域科研产出的社会影响力走势如图 6 所示。

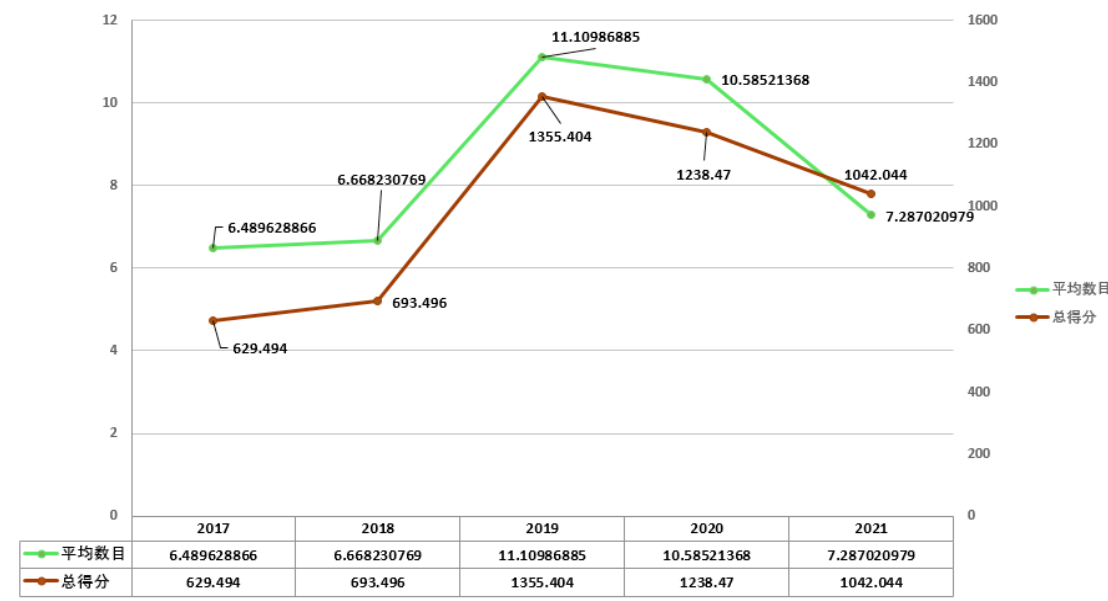


图 6 主题演化桑基图

由图 6 可知,自 2017 年信息素养相关研究的社会影响力总 Altmetric 得分呈上升趋势,2019 年总 Altmetric 得分最高,为 1 355.404,近两年 2020、2021 年社会影响力得分稍有下降但也保持在 1 000 以上。总体上来看,近五年信息素养领域相关研究的社会影响力分值呈现上升势头。与此同时,就篇均社会影响力来看,近五年每年文献量都在 200 篇附近,篇均社会影响力分值都在 10 附近,而每年的总社会影响力分值差异较大,即存在超高社会影响力分值的文献,即对应可以理解为学术影响力指标中的高被引论文、热点论文^[17]等。由于社会影响力分值是随着时间不断变化的,所

以参考高被引论文的计算方法,选取总文献篇数的前 1% 的 altmetric 值作为高社会影响力的阈值,即社会影响力得分前 11 的为高社会影响力文献,这些文献题录详情如表 2 所示。具体来看,近五年信息素养主题社会影响力最高的文章为 2019 年的 *Approaching literature review for academic purposes: The Literature Review Checklist* 是一篇学术素养综述文献,其次 是 2021 年 *Improving college students fact-checking strategies through lateral reading instruction in a general education civics course*,探讨的是提高大学生群体虚假信息甄别的策略方法。

表 2 各年代区间内主题高频词分布表

年份	ALT值	标题	来源出版物
2017	61.922	Glaring Chemical Errors Persist for Years on Wikipedia	Journal of Chemical Education
2018	73.026	The Practice and Promise of Critical Information Literacy: Academic Librarians' Involvement in Critical Library Instruction	College & Research Libraries
	250.9	Approaching literature review for academic purposes: The Literature Review Checklist	Clinics
	116.6	Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective	Behaviour & Information Technology
2019	83.13	Undergraduate students' experiences of using information at the career fair: A phenomenographic study conducted by the libraries and career center	Journal of Academic Librarianship
	81.23	Effects of the flipped classroom instructional strategy on students' learning outcomes: a meta-analysis	Educational Technology Research & Development
	79.55	Critical librarianship in health sciences libraries: an introduction	Journal of the Medical Library Association
	187.01	Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools	Psychological Science in the Public Interest
2020	184.85	Educating Youth for Nonexistent/Not Yet Existing Professions	Educational Policy
	109.864	Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation	Journal of Medical Internet Research
	217.35	Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course	Cognitive Research: Principles and Implications
2021	98.98	Students Evaluating and Corroborating Digital News	Scandinavian Journal of Educational Research

5 总结

近五年关于信息素养领域的相关科研成果在数量上呈现上升趋势,通过划分时间区间窗口,基于 LDA 模型编写 python 程序实施主题演化分析发现,不同时间窗口下的研究主题连续性较强,不同主题之间的相似度较高。近五年信息素养领域相关研究主要集中于信息素养教育教学、学术图书馆建设、数据(数字)素养、图书馆员能力提升、社会信息服务等方面,并且随着时事的发展,近两年,在线教育、健康素养、虚假信息识别等信息素养相关的研究主题也不断引起业界关注,与此同时,社会生活中出现的信息匮乏问题,也逐步演化成为信息难民现象。

本文的局限之处在于,本文的数据集是基

于近五年信息素养 SCI 和 SSCI 中的文献数据,未将中文信息素养相关研究进行统一分析,所以相关研究结论只局限于本数据集中的信息素养相关研究。另外,在主题演化分析中,通过 LDA 模型析出的各个年代的主题名,为笔者依据主题可视化、高频词分布等总结提取命名,可能存在因专业知识的欠缺导致归纳总结的主题名不规范的问题。

参考文献

- [1] 姜颖,陆广琦.国外高校信息素养教育规划分析与启示[J].图书馆建设,2021(5): 162-170+176.
- [2] 于良芝,王俊丽.从普适技能到嵌入实践——国外信息素养理论与实践回顾[J].中国图书馆学报,2020,46(2): 38-55.
- [3] CILIP: what is literacy information? [EB/OL]. [2022-11-11]. <https://www.cilip.org.uk/news/news.asp?id=421972>

- [4] 中国图书馆学会“十四五”发展规划纲要(2021-2025年)[EB/OL]. [2022-11-11]. <http://www.lsc.org.cn/contents/1299/15358.html>
- [5] Altmetric. Sources of Attention[EB/OL]. [2022-11-11]. <https://help.altmetric.com/support/solutions>
- [6] (美)安德鲁·凯莱赫, 亚当·凯莱赫著. 陈子墨, 刘瀚文译. 机器学习实践[M]. 北京: 机械工业出版社, 2020: 126.
- [7] 冯兴东著. 分布式统计计算[M]. 上海: 上海财经大学出版社, 2018: 74.
- [8] LDA 主题的建模过程学习[EB/OL]. [2022-11-11]. <https://www.cnblogs.com/luckyplj/p/12718922.html>
- [9] Tazibt A A, Aoughlis F. Latent Dirichlet allocation-based temporal summarization[J]. International Journal of Web Information Systems, 2019, 15(1): 83-102.
- [10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.
- [11] LDA 主题个数的确定[EB/OL]. [2022-11-11]. https://blog.csdn.net/weixin_31468621/article/details/112195887
- [12] 周健, 张杰, 屈冉, 闫石. 基于 LDA 的国内外区块链主题挖掘与演化分析[J]. 情报杂志, 2021, 40(9): 161-169.
- [13] 王博雅. Schools 联盟 iCaucus 成员高校图书馆数字素养教育调查研究[J]. 图书馆工作与研究, 2021(4): 29-36.
- [14] 杨思洛著. 替代计量学理论、方法与应用[M]. 北京: 科学出版社, 2018: 74.
- [15] 邱均平, 张心源, 董克. Altmetrics 指标在机构知识库中的应用研究[J]. 图书情报工作, 2015, 59(2): 100-105.
- [16] PIWOWAR H. Altmetrics: Value all research products[J]. Nature, 2013, 493(7431):159.
- [17] Citation Threshold [EB/OL]. [2022-11-11]. <http://esi.help.clarivate.com/Content/citation-thresholds.htm>