学科主题视角下的科学论文被引影响因素差异性研究*
——以图书情报学科为例

傅柱　张倩　刘鹏
（江苏科技大学经济管理学院，镇江 212100）

摘要：从学科主题角度探究被引频次影响因素的差异，为科研人员撰写和优化论文提供有针对性的参考，也为科研评价研究提供新的视角和思路。以18种图书情报学CSSCI收录期刊在2011—2020年发表的43 228篇有效论文为样本，运用LDA主题模型对论文摘要进行主题提取和识别，从总体、主题、主题分类3个角度采用负二项回归模型对年均被引频次的影响因素进行实证研究。研究表明，学科主题间年均被引频次的影响因素存在共性和差异性。基金资助、论文篇幅、论文年龄、下载量对年均被引频次的影响具有一致性，标题长度、关键词数量、作者数量等因素对年均被引频次的影响呈现出差异性。

关键词：被引频次；主题分类；回归分析；影响因素；图书情报学
中图分类号：G353.1　DOI：10.3772/j.issn.1673-2286.2024.09.002
引文格式：傅柱，张倩，刘鹏. 学科主题视角下的科学论文被引影响因素差异性研究：以图书情报学科为例[J]. 数字图书馆论坛，2024，20（9）：16-26.

学术论文的被引频次指标是指论文正式发表之后被其他发表文献引用的累积频次，被引频次作为学术论文的重要定量指标，在一定程度上可以反映论文质量，利用被引频次来评判论文的质量也是最为常见的方式[1]。一方面，大量的引用总是集中于少量的论文，80%的引用来自20%的高被引论文；另一方面，大部分论文的被引频次很少，甚至还有不少零被引论文。这一偏态分布现象已成为学界的共识[2]。被引频次是评判论文质量的重要因素，质量越高的论文越能够得到同行的认可，被引频次相应更高。除论文自身质量以外，论文被引频次还受到外在因素的影响。论文质量难以准确测量，但影响论文被引频次的外在因素的测量具有现实的可行性。目前，学者们对论文被引频次外部影响因素的研究较为成熟，但主要局限在单一学科的视角，未考虑学科内部主题不同是否会使被引频次影响因素呈现不同的结果。为此，本文从学科主题角度进行研究，以期更加细致地了解不同主题研究成果的被引用情况。希望能够为学术评价中的被引频次影响因素研究提供新的视角，为研究者实施科研活动和撰写高水平论文提供参考。

1　相关研究综述

研究者们将文献被引频次的影响因素归纳为论文因素、作者因素、期刊因素，并针对这3类因素对被引频次的影响展开大量研究。论文因素能够直接影响被

收稿日期：2024-05-20
*本研究得到国家社会科学基金青年项目“面向AI4S的场景化智慧知识服务框架研究”（编号：24CTQ030）资助。

引频次，常见的影响因素有参考文献、标题、论文篇幅等。①参考文献影响被引频次的研究。Antoniou等[3]通过单变量和多元线性回归模型评估了参考文献的数量对被引频次的影响，结果表明参考文献的数量越多被引频次也越多。Roth等[4]提出了一种基于参考文献结构的引文预测方法，指出了参考文献越新，论文越能够获得更多的被引频次。②标题影响被引频次的研究。Jacques等[5]研究标题与被引之间的关系，结果表明标题的字数和标题的结构对被引率有影响。Rossi等[6]指出，为了最大限度地提高论文的影响力，作者应有针对性地选择准确且简洁的标题。③论文篇幅影响被引频次的研究。张振伟等[7]通过研究指出，论文的版面数越多，论文的被引频次越多，此外是否为重点专题论文以及论文类型和学科分类也可能与被引频次有关联。

作者因素也会对被引频次产生影响，具体包括作者数量、作者间合作关系、作者所属机构等。Leimu等[8]在研究生态学被引频次的影响因素时指出，作者数量对被引频次有影响。Borsuk等[9]认为被引频次与作者的性别无关。论文的被引频次还与作者的声望以及作者早期的被引频次有关，作者的声望和地位越高，被引频次就越高。学者普遍认为合作网络能够影响作者获取信息的能力，从而影响论文的影响力。部分学者对此进行了实证研究，如杜建等[10]对医学领域不同学科作者的合作度与论文影响力之间的关系进行研究，发现多作者、国际、机构合作的论文被引频次显著高于单作者、国内和机构内合作论文。王崇锋等[11]探讨了合作网络与知识网络的中心性特征与结构特征对被引频次的影响，合作网络中心性特征对被引频次有显著的倒U型影响。

期刊因素对被引频次的影响效果也十分显著，发表在高影响力、高等级期刊上的论文更容易得到关注，有很大概率能够成为高被引论文，这已经基本成为共识。此外，随着期刊出版模式的发展和改变，论文是否开放获取也会影响论文的受关注度和被引频次。其中，研究者较多关注期刊影响因子对被引频次的影响。研究人员倾向于将研究成果发表在影响因子高的期刊上，以此获得更多的关注与引用。杨莉等[12]在对被引频次的预测研究中加入了期刊影响因子指标进行预测。

对被引频次影响因素的研究已经比较全面，在论文层面，大多数研究从标题长度、关键词的数量、参考文献的数量、基金资助、论文的篇幅等角度展开；在作者层面，主要从作者的数量、作者的年龄、作者之间的合作关系、第一作者的发文量等角度展开研究；在期刊层面，一般考虑期刊影响因子、期刊总发文量等。也有少数学者会考虑到论文的主题特征，如主题排名、主题规模、主题的多样性等[13]。当融入主题因素进行评价研究时，能更为精确地反映主题内部的影响力情况。国外学者研究证明，主题对被引频次的影响十分显著，热点主题往往会吸引更多的引用[14]。目前，随着数据挖掘技术的不断成熟与发展，国内外对主题的研究也相对成熟，但是从学科主题角度对被引频次影响因素的研究还不够普及[15-16]。属于同一学科领域的文献，由于研究主题方向不同，受关注度也有差异。因此，本研究从主题视角多层面研究各因素对被引频次的影响，揭示不同主题下的共性影响因素和差异性影响因素，为科研评价提供重要参考。

2　主题抽取与文献分类

2.1　数据采集与处理

（1）数据采集。研究数据来源于中国知网，通过高级检索依次在文献来源中输入期刊名称，主要选取图书情报领域的18种CSSCI期刊，包括《大学图书馆学报》《国家图书馆学刊》《情报科学》《情报理论与实践》《情报学报》《情报杂志》《情报资料工作》《图书馆》《图书馆工作与研究》《图书馆建设》《图书馆论坛》《中国图书馆学报》《图书馆学研究》《图书馆杂志》《图书情报工作》《图书情报知识》《图书与情报》《数据分析与知识发现》。选定在2011年1月1日—2020年12月31日发表的期刊文献，共检索获得47 034篇文献，以来源库、题名、作者、单位、文献来源、关键词、摘要、发表时间、基金、年、页码等为自定义字段导出题录数据。研究框架如图1所示。

缺失的数据通过Python程序爬虫获取，以年均被引频次为因变量，以论文因素（标题长度、基金资助、关键词数量、论文年龄、论文篇幅、中文参考文献占比、参考文献数量、下载量），作者因素（作者数量、第一作者发文量、第一作者平均下载量、第一作者机构、跨单位合作数量），期刊因素（期刊影响因子）为自变量获取相关数据。具体变量及定义如表1所示。

（2）数据处理。去除征稿启事、选题指南、序言，以及无摘要、无作者等无效文献，共获得有效文献43 228

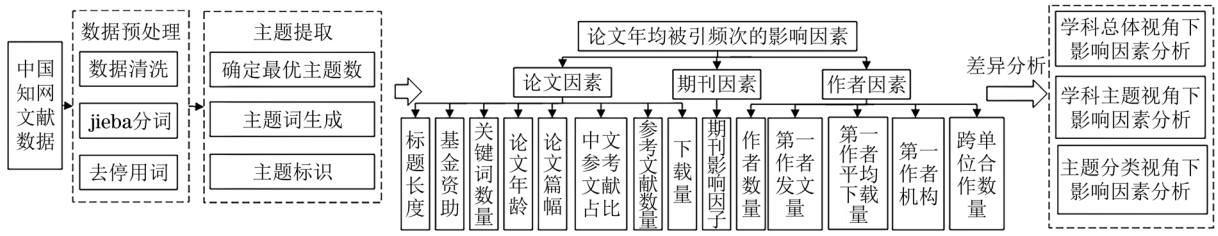

图1　研究框架

表1　变量选取和定义

类别	变量名称	变量定义
作者因素	作者数量	文章的作者数量
	第一作者发文量	文章第一作者总发文量
	第一作者平均下载量	文章第一作者总下载量与发文量之比
	第一作者机构	第一作者所属机构类型，1=其他机构，2=高校
	跨单位合作数量	文章合作单位数量
期刊因素	期刊影响因子	文章来源期刊影响因子
论文因素	标题长度	文章标题长度，包括中英文字符总数
	关键词数量	文章关键词数量
	论文年龄	文章发表至检索时间的年份差值
	论文篇幅	文章的页数
	中文参考文献占比	文章中文参考文献与总参考文献数量的比值
	参考文献数量	文章参考文献的数量
	下载量	文章的下载量
	基金资助	0=未获得基金资助，1=获得基金资助
因变量	年均被引频次	文章总被引频次与论文年龄之比

篇。汇总摘要作为语料库，运用jieba分词工具对原始语料进行分词、去停用词。

2.2　LDA主题提取

利用LDA主题模型[17]对摘要语料库进行主题提取，在主题模型训练之前需要预先设定主题数量。困惑度和余弦相似度是目前自然语言处理中常用的评价指标。选取不同的主题数量，计算主题间的平均余弦相似度和困惑度，困惑度的得分越低，说明模型的效果越好。平均余弦相似度越小，主题的结构越稳定。困惑度和余弦相似度的结果如图2和图3所示。

结合困惑度和余弦相似度的结果，确定最优的主题数量为8个。使用sklearn库对数据进行主题建模，LDA算

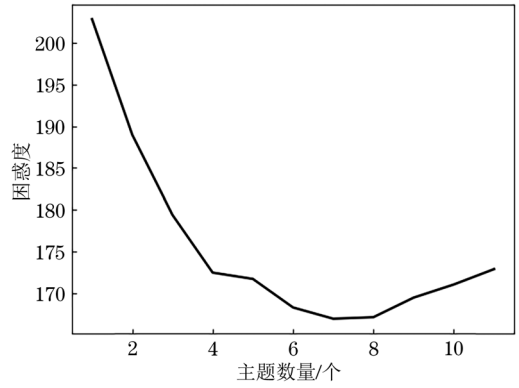

图2　主题困惑度曲线

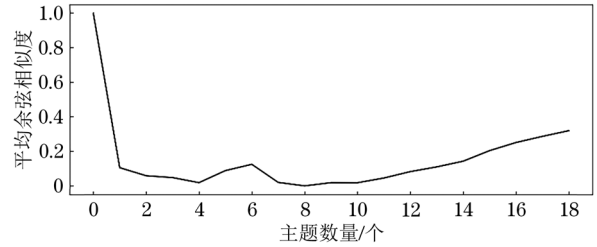

图3　平均余弦相似度曲线

法的参数设置为α=0.1，β=0.01，主题数量设置为8，迭代次数为1 000。根据LDA模型生成的各研究主题，人工确定8个研究主题标签。LDA模型主题提取结果如表2所示。

2.3　文献分类

利用LDA主题模型输出的主题概率分布，将每篇文献分配到最相关的主题下，文献主题分类结果如表3所示。

3　实验结果与分析

3.1　描述统计和相关分析

（1）描述性统计分析。对自变量和因变量进行描

表2　LDA模型主题提取结果

主题类别	Top 20主题词
学术评价与文献计量	评价、分析、文献、领域、方法、专利、学术、指标、期刊、论文、学科、发展、图书馆学、情报学、图书、作者、合作、理论、主题、影响力
图书馆服务	图书馆、服务、发展、文化、高校、读者、建设、模式、社会、空间、分析、方面、活动、文章、问题、工作、内容、事业、合作、制度
网络舆情	网络、舆情、传播、事件、社会、研究、分析、应急、古籍、危机、突发事件、过程、编目、书目、文献、政务、藏书、话题、问题、历史
文本挖掘	方法、模型、数据、主题、结果、语义、分析、文本、特征、分类、算法、领域、检索、技术、关系、实验、研究、利用、本体、结论
用户信息行为	用户、信息、影响、研究、因素、模型、移动、分析、理论、网络、方法、过程、社交、媒体、结果、社区、平台、结论、关系、公众
图书馆建设	资源、数字、图书馆、建设、文献、服务、信息、系统、分析、问题、平台、智库、高校、机构、方面、网站、技术、利用、数据库、文章
企业和政府情报分析	信息、情报、企业、分析、政府、技术、发展、过程、政策、管理、研究、竞争、理论、方法、国家、机制、协同、结果、体系、基础
信息教育和科研素养	数据、教育、学科、信息、高校、科研、素养、科学、馆员、分析、专业、能力、课程、团队、研究、教学、方面、数据管理、学生、调查

表3　LDA模型主题文献分类（示例）

文献标题	主题类别	主题概率
儒法两家经典的共词分析与研究	文本挖掘	[0.222 000 36，0.004 386 27，0.206 234 67，0.549 832 65，0.004 386 55，0.004 386 89，0.004 386 20，0.004 386 41]
基于文献调研的国内外高校信息素养教学内容与模式趋势探析	信息教育和科研素养	[0.347 980 69，0.007 246 93，0.007 247 17，0.007 247 66，0.007 248 51，0.007 247 62，0.007 247 96，0.608 533 47]
新信息环境下高校信息检索课教学方式的优化策略	信息教育和科研素养	[0.002 873 73，0.002 874 81，0.002 874 72，0.002 874 15，0.002 874 34，0.138 440 18，0.002 874 01，0.844 314 06]
北京地区高校学生图书馆焦虑测量分析——基于图书馆焦虑量表的修订与验证	用户信息行为	[0.003 356 55，0.292 208 73，0.003 356 06，0.003 357 37，0.402 091 01，0.003 355 98，0.003 356 23，0.288 918 07]
人文社科领域科学数据使用特征分析——基于《中国社会科学》样本论文的实证研究	学术评价与文献计量	[0.544 495 95，0.047 083 65，0.003 247 29，0.121 346 09，0.003 247 22，0.003 247 39，0.003 247 34，0.274 085 07]

述性统计，统计结果如表4和表5所示。第一作者发文量、第一作者平均下载量、下载量标准差较大，呈现离散分布。标题长度均值为19.92，作者数量均值为2.03，

表4　连续变量描述性统计结果

类别	变量名称	论文数量/篇	均值	标准差	最小值	最大值
作者因素	作者数量	43 228	2.03	1.12	1.00	20.00
	第一作者发文量	43 228	54.14	254.00	1.00	679.00
	第一作者平均下载量	43 228	624.20	3 405.67	3.18	36 222.00
	跨单位合作数量	43 228	1.41	0.68	1.00	11.00
期刊因素	期刊影响因子	43 228	3.26	0.93	1.95	9.29
论文因素	标题长度	43 228	19.92	6.24	1.00	61.00
	关键词数量	43 228	4.03	0.99	0.00	15.00
	论文年龄	43 228	8.02	3.20	3.00	13.00
	论文篇幅	43 228	6.07	2.41	1.00	30.00
	中文参考文献占比	43 228	0.70	0.29	0.00	1.00
	参考文献数量	43 228	11.12	9.74	0.00	165.00
	下载量	43 228	752.55	930.96	0.00	42 308.00
因变量	年均被引频次	43 228	2.34	3.48	0.00	83.00

表5　分类变量描述性统计结果

变量名称	类别	论文数量/篇	比例/%
第一作者机构	“双一流”高校	17 622	40.77
	普通高校	18 436	42.64
	其他机构	7 170	16.59
基金资助	国家级	15 466	35.78
	省部级	5 038	11.65
	其他基金	4 012	9.28
	无基金	18 712	43.29

跨单位合作数量均值为1.41，期刊影响因子均值为3.26，关键词数量均值为4.03，论文年龄均值为8.02，论文篇幅均值为6.07，中文参考文献占比均值为0.70，参考文献数量均值为11.12。第一作者机构大多为“双一流”高校以及普通高校，占比分别为40.77%、42.64%。受国家级基金资助的论文占比35.78%。

（2）多重共线性检验。解释变量间的多重共线性会对回归的结果产生影响。为了避免由变量之间显

著相关性导致的多重共线性问题，在回归分析之前需要对各变量之间相关系数进行检验。采用Spearman相关系数对变量之间的相关性进行检验，如表6所示。可以看出，大部分相关系数较小，进一步采用方差膨胀因子（Variance Inflation Factor，VIF）进行检验，VIF值均小于5。因此，各变量之间不存在严重的共线性问题，可以将数据导入回归模型进行实证检验。

表6　Spearman相关系数检验结果

变量名称	作者数量	第一作者发文量	第一作者平均下载量	跨单位合作数量	期刊影响因子	标题长度	关键词数量	论文年龄	论文篇幅	中文参考文献占比	参考文献数量	下载量
作者数量	1.000											
第一作者发文量	0.164**	1.000										
第一作者平均下载量	0.323**	0.242**	1.000									
跨单位合作数量	0.461**	0.117**	0.018**	1.000								
期刊影响因子	0.198**	0.119**	0.191**	0.122**	1.000							
标题长度	0.110**	0.010**	0.100**	0.050**	0.047**	1.000						
关键词数量	0.151**	0.089**	0.162**	0.105**	0.114**	0.145**	1.000					
论文年龄	−0.162**	−0.019**	−0.206**	−0.117**	−0.029**	−0.153**	−0.165**	1.000				
论文篇幅	0.163**	0.058**	0.197**	0.116**	0.088**	0.078**	0.153**	−0.504**	1.000			
中文参考文献占比	−0.059**	−0.038**	−0.071**	−0.046**	−0.043**	−0.006	−0.055**	0.124**	−0.192**	1.000		
参考文献数量	0.151**	0.046**	0.148**	0.098**	0.166**	0.079**	0.114**	−0.382**	0.517**	−0.053**	1.000	
下载量	0.121**	0.035**	0.131**	0.069**	0.153**	0.057**	0.091**	−0.273**	0.410**	−0.057**	0.345**	1.000
VIF	1.290	1.005	1.002	1.249	1.073	1.037	1.062	1.342	1.581	1.026	1.357	1.096

注：分类变量未放入表中；**表示在5%水平（双边）下通过显著性检验。

3.2　学科总体视角下年均被引频次影响因素分析

年均被引频次密度分布图（见图4）显示，年均被引频次呈左偏态分布，这意味着大部分数据点集中在数值较低的一端，而数值较高的数据较少。

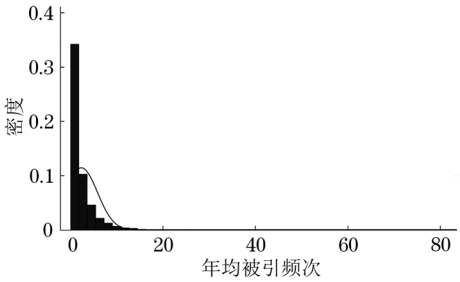

图4　年均被引频次密度分布图

由于年均被引频次呈现偏态分布，传统的多元线性回归模型并不合适。年均被引频次属于离散型变量，取值是典型的计数型，因此考虑计数型回归分析。为了选择合适的回归模型，使用Stata 18.0软件[18]的泊松回归、负二项回归、零膨胀泊松回归、零膨胀负二项回归进行了回归分析。研究发现，样本的方差大于均值，可能存在过度离散现象，不符合泊松回归模型的要求，因此选择负二项回归模型。使用负二项回归和零膨胀负二项回归进行数据分析与验证，采用赤池信息准则（Akaike Information Criterion，AIC）、贝叶斯信息准则（Bayesian Information Criterion，BIC）进行模型的检验。AIC与BIC相类似，是衡量统计模型拟合优良性的准则，定义式如式（1）和式（2）所示。

$$V_{AIC} = -2l + 2p \quad (1)$$

$$V_{BIC} = -2l + p\ln n \quad (2)$$

式中：l为对数似然值，p为模型的参数数量，n为观测样本数量。AIC、BIC值越小，模型拟合的效果越好。AIC、BIC值拟合结果如表7所示，负二项回归的AIC、BIC值略低于零膨胀负二项回归，最终选择的回归模型为负二项回归。

负二项回归的似然比统计量为35 540.37，表明模型与零假设之间存在显著差异，即模型整体是显著的；R^2为0.201 8，表明模型的拟合效果可以接受。总体视角下年均被引频次影响因素的负二项回归分析结果如表8所示，z为模型统计量，P小于0.1则变量的影响显著。大多数解释变

量呈现显著影响，说明整体回归效果比较好。在1%水平下，作者数量、第一作者平均下载量、基金资助、期刊影响因子、论文篇幅、中文参考文献占比以及下载量均对论文年均被引频次产生显著正向影响，第一作者发文量在10%水平下具有显著正向作用；论文年龄在1%水平下对年均被引频次具有显著的负向作用；标题长度、第一作者机构、跨单位合作数量、关键词数量、参考文献数量在1%、5%、10%的水平（双边）下均未通过显著性检验。

表7　AIC、BIC值拟合结果

模型	观测样本数量	空模型的对数似然值	模型的对数似然值	自由度	AIC值	BIC值
负二项回归	43 228	–88 061.86	–70 291.68	16	140 615.4	140 754.1
零膨胀负二项回归	43 228	–88 061.86	–70 291.68	18	140 619.4	140 775.5

表8　学科总体视角下年均被引频次影响因素的负二项回归分析结果

类别	变量名称	系数	标准误差	*z*	*P*
作者因素	作者数量	0.015 115 3	0.003 723 3	4.06	<0.001
	第一作者发文量	0.000 014 5	0.000 008 3	1.76	0.078
	第一作者平均下载量	0.000 003 2	0.000 000 6	5.28	<0.001
	第一作者机构	–0.000 568 7	0.011 218 3	–0.05	0.960
	跨单位合作数量	0.007 995 6	0.005 765 2	1.39	0.165
期刊因素	期刊影响因子	0.021 478 3	0.004 723 3	4.55	<0.001
论文因素	标题长度	0.000 947 5	0.000 596 7	1.59	0.112
	基金资助	0.118 954 8	0.008 243 4	14.43	<0.001
	关键词数量	–0.000 881 4	0.003 840 9	–0.23	0.818
	论文年龄	–0.103 072 4	0.001 425 4	–72.31	<0.001
	论文篇幅	0.024 425 5	0.002 039 2	11.98	<0.001
	中文参考文献占比	0.167 481 9	0.014 463 8	11.58	<0.001
	参考文献数量	0.000 624 5	0.000 432 0	1.45	0.148
	下载量	0.000 703 7	0.000 011 5	61.37	<0.001
	截距	0.349 381 5	0.034 288 1	10.19	<0.001

3.3　学科主题视角下年均被引频次影响因素分析

受篇幅所限，对8个学科主题进行负二项回归的过程不一一列出，各个回归模型的简略汇总结果如表9所示。

从表9可以看出，不同主题下年均被引频次的影响因素存在一定差异。

（1）作者因素。总体回归结果显示，年均被引频次与作者数量、第一作者发文量、第一作者平均下载量显著相关。然而，学科主题回归的发现与之不同。①作者数量对年均被引频次的显著正向作用只体现在学术评价与文献计量、网络舆情、文本挖掘3个主题中。在学术评价与文献计量主题下，较多的作者数量可能反映了团队合作，学术评价领域注重学术研究成果的评估、测量与分析，多作者合作可能提升了研究的可信度和影响力；网络舆情主题通常涉及公众关注度较高的话题，作者数量多，论文可能反映了多方面的专业意见和观点，能吸引更多的引用；文本挖掘主题涉及复杂的文本数据，相关研究需要全面和深入的数据分析，多位作者之间的合作可以加强数据的多维度分析，也有助于提出更具有深度和广度的解决方案，进而吸引更多的引用。②第一作者发文量在各主题下均未通过显著性检验，说明第一作者发文量对论文被引频次没有显著影响。③第一作者平均下载量对年均被引频次的显著正向作用体现在图书馆服务、信息教育和科研素养、企业和政府情报分析3个主题中。第一作者的平均下载量较高，表明该作者的论文受到更多的关注，这对该主题论文学术影响力的提高有帮助。虽然第一作者机构、跨单位合作数量在总体回归中没有体现出显著作用，但是在文本挖掘主题中第一作者机构对年均被引频次具有正向作用，企业和政府情报分析主题中第一作者机构对

表9　学科主题视角下年均被引频次影响因素的负二项回归分析结果

类别	变量名称	学术评价与文献计量	图书馆服务	网络舆情	文本挖掘	用户信息行为	图书馆建设	企业和政府情报分析	信息教育和科研素养
作者因素	作者数量	+		+++	++				
	第一作者发文量								
	第一作者平均下载量		+++					+++	+
	第一作者机构				+			-	
	跨单位合作数量								+
期刊因素	期刊影响因子			+++		+++			
论文因素	标题长度			+					
	基金资助	+++	+++	+++	+++	+++	+++	+++	+++
	关键词数量								+
	论文年龄	---	---	---	---	---	---	---	---
	论文篇幅	+++	+++	+++	+++	+++	+++	+++	+++
	中文参考文献占比	+++	+++	+++		++	+++		+++
	参考文献数量	-							++
	下载量	+++	+++	+++	+++	+++	+++	+++	+++

注：+++、++、+分别表示在1%、5%、10%的水平下通过显著性检验，并且系数符号为正；---、--、-分别表示在1%、5%、10%的水平下通过显著性检验，并且系数符号为负；空白表示未通过显著性检验。

年均被引频次具有负向作用，而跨单位合作数量仅在信息教育和科研素养主题中对年均被引频次具有正向作用。在信息教育和科研素养主题中，通过跨单位合作能够结合不同领域和学科的专业知识和研究资源，提升研究的多样性和深度，从而提升研究的创新性和学术质量。

（2）期刊因素。在网络舆情、用户信息行为2个主题下期刊影响因子在1%的水平下通过显著性检验，表明期刊影响因子在这2个主题下对年均被引频次具有正向作用，但其影响在其他主题下未通过显著性检验。

（3）论文因素。基金资助、论文年龄、论文篇幅、下载量的分主题回归结果与总体回归结果保持一致。中文参考文献占比仅在文本挖掘、企业和政府情报分析2个主题下对年均被引频次没有显著影响，在其他6个主题中对年均被引频次均有显著正向作用。标题长度仅在网络舆情主题中具有显著正向作用。网络舆情主题涉及公众关注和信息传播，较长的标题可能更能够吸引注意，长标题倾向于包含更多的关键词，这些关键词与当时的热门话题相关，进而增加被引用的可能性。参考文献数量在学术评价与文献计量主题下对年均被引频次有显著负向作用，在信息教育和科研素养主题下具有显著正向作用。在学术评价与文献计量主题下，研究往往涉及大量的文献引用和计量分析，引用的大量参考文献未必对新的研究有直接的贡献，从而影响了学术影响力。信息教育和科研素养主题涉及科研实践等方面，引用的文献有助于建立坚实的理论和知识基础，体现了研究者对该领域主题全面和深入的理解，能增强研究的可信度和引用价值。

通过对不同主题下年均被引频次影响因素的分析，可以看出学术研究在不同主题下表现出显著的差异，这些差异反映了各主题研究特定的引用习惯，解释了不同因素在提升论文学术影响力方面的具体作用。通过深入理解这些因素在特定主题中的作用机制，能够帮助学者提升学术影响力。

3.4　主题热度分区视角下年均被引频次影响因素分析

学者们往往以发文量和引文量为考察学科主题影响力的基本指标[19-20]。然而，不同主题文献的外部特征、研究内容等具有较大差异。因此，进一步探究不同主题热度下年均被引频次影响因素的差异情况。参考李秀霞等[21]的研究，统计各研究主题的逐年累计引文量和发文量。设有R个主题，统计某年某个主题的引文量与发文量之比E_r（r=1，…，R），某年所有主题对应文献引文量与发文量之比D_r，根据E_r、D_r计算引文等级q_{rt}，如式（3）~式（5）所示。

$$E_r = \frac{\sum_{i=1}^{N_r} C_i}{N_r} \quad (3)$$

$$D_r = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} C_{ir}}{N} \quad (4)$$

$$q_{rt} = \frac{E_r - D_r}{\sqrt{\frac{\sum_{r=1}^{R}(E_r - D_r)^2}{R}}} \quad (5)$$

式中：C_i表示第i篇论文的引文量，N_r表示某年第r个主题发文量；C_{ir}表示第r个主题、第i篇论文的引文量，N表示某年所有主题发文量；t表示时间段（t=1，2，3，……）。

某研究主题在某年的发文量记为N_{rt}，对q_{rt}、N_{rt}的值与时间段进行Spearman相关系数分析，得到与时间段的相关系数，建立战略坐标系。发文呈现递增趋势、引文也呈现递增趋势则划分在热门分区，发文呈递减趋势、引文呈现递增趋势划分为潜力分区，发文呈递减趋势、引文呈现递减趋势划分为衰退分区，发文呈递增趋势、引文呈现递减趋势划分为冷门分区。将各主题划分到不同的分区，探究不同热度分区下年均被引频次影响因素是否存在一定差异。图书情报学主题热度分区图如图5所示，将8个主题分类到各分区下，横轴表示的是发文趋势，纵轴表示的是引文趋势。第一象限发文量大、被引频次高，代表热门分区，包括用户信息行为主题、文本挖掘主题；第二象限发文量小但被引频次高，代表潜力分区，包括企业和政府情报分析主题；第三象限发文量小、被引频次低，代表衰退分区，包括网络舆情主题、图书馆服务主题、图书馆建设主题、学术评价与文献计量主题、信息教育和科研素养主题；第四象限发文量大、被引频次低，代表冷门分区，目前无主题划分，说明图书情报学科的主要研究主题均处于较高热度。

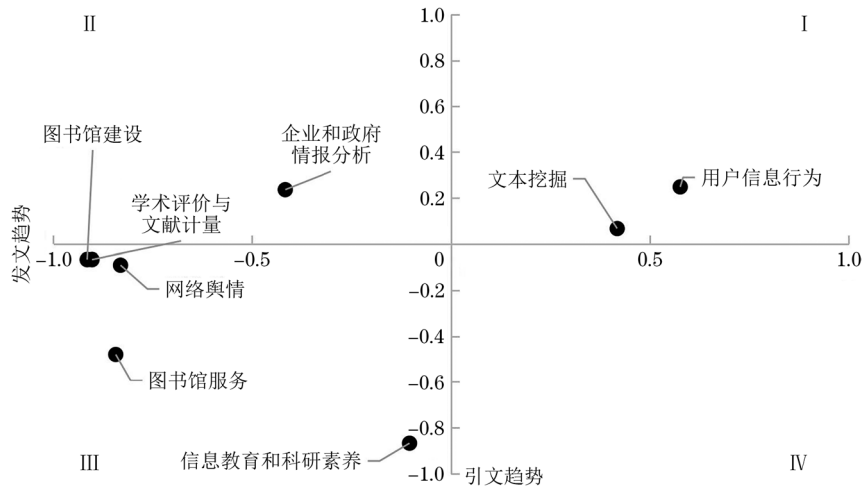

图5　图书情报学主题热度分区图

主题热度分区下年均被引频次影响因素的负二项回归分析结果如表10所示。从作者角度看，作者数量在热门和衰退分区下对年均被引频次均具有显著正向作用，第一作者平均下载量在潜力和衰退分区下对年均被引频次具有正向作用，第一作者机构在潜力分区中对年均被引频次具有显著负向作用，第一作者发文量、跨单位合作数量对年均被引频次没有显著影响。从期刊角度看，期刊影响因子仅在衰退分区下对年均被引频次具有正向作用。衰退分区下，影响因子高的期刊能够显著提升论文的关注度，衰退分区的发文趋势和引文趋势均在减弱，受关注较少，而在高影响因子期刊上发表能够增加被引用的机会。从论文角度看，基金资助、论文篇幅、下载量在3个分区都具有正向作用，但标题长度仅对衰退分区论文年均被引频次具有正向作用。长标题通常能清楚地描述研究内容，从而吸引更多的引用，尤其在衰退分区下，长标题更为重要。中文参考文献占比在热门分区和衰退分区中均具有正向作用，关键词数量在各分区中均未通过显著性检验。此外，论文年龄在各分区中均有显著负向影响。基于各影响因素在不同分区的差异，学者可在不同分区主题下制定有效的研究和发表策略。

3.5　稳健性分析

为了验证不同主题下年均被引频次影响因素的稳

表10　主题热度分区下年均被引频次影响因素的负二项回归分析结果

类别	变量名称	热门	潜力	衰退
作者因素	作者数量	+++		+
	第一作者发文量			
	第一作者平均下载量		+++	+++
	第一作者机构		-	
	跨单位合作数量			
期刊因素	期刊影响因子			+++
论文因素	标题长度			+
	基金资助	+++	+++	+++
	关键词数量			
	论文年龄	---	---	---
	论文篇幅	+++	+++	+++
	中文参考文献占比	++		+++
	参考文献数量			
	下载量	+++	+++	+++

注：+++、++、+分别表示在1%、5%、10%的水平下通过显著性检验，并且系数符号为正；---、--、-分别表示在1%、5%、10%的水平下通过显著性检验，并且系数符号为负；空白表示未通过显著性检验。

定性，进一步采用零膨胀负二项回归模型进行稳健性检验。由负二项回归模型和零膨胀负二项回归模型的结果可知，两个模型的参数估计和显著性水平具有一致性，在主要影响因素如标题长度、作者数量、第一作者平均下载量、跨单位合作数量、期刊影响因子和中文参考文献占比等变量上结果是一致的。将不同主题的数据分成2011—2015年、2016—2020年两组分别进行负二项回归，发现基金资助、论文年龄、论文篇幅、下载量对年均被引频次的影响均保持一致。由于时间段的划分，标题长度、关键词数量、作者数量等因素的作用存在差异。

4　结论与启示

以图书情报学18种CSSCI收录期刊为研究对象，采集了2011—2020年的共43 228篇有效论文相关数据。采用LDA主题模型和负二项回归方法从学科主题的角度探究论文外部特征与年均被引频次的影响关系，得出以下结论。

（1）从学科主题角度看，不同主题下年均被引频次的影响因素存在显著差异。作者数量在学术评价与文献计量、网络舆情、文本挖掘3个主题中显著正向影响年均被引频次；第一作者发文量在各主题中均无显著影响；第一作者平均下载量仅在图书馆服务、企业和政府情报分析、信息教育和科研素养主题下有显著正向作用；第一作者机构在文本挖掘主题下有正向作用，在企业和政府情报分析主题下有负向作用；跨单位合作数量在信息教育和科研素养主题中有正向作用；期刊影响因子在网络舆情和用户信息行为主题中有显著正向作用；中文参考文献占比在多数主题中有显著正向作用；标题长度在网络舆情主题中有显著正向作用；参考文献数量在学术评价与文献计量主题中有显著负向作用，在信息教育和科研素养主题中有显著正向作用。

（2）从主题分类角度看，图书情报学研究主题分为热门、潜力和衰退3类。热门分区包含用户信息行为、文本挖掘2个主题；潜力分区包括企业和政府情报分析1个主题；衰退分区包括网络舆情、图书馆服务、图书馆建设、学术评价与文献计量、信息教育和科研素养5个主题。各分区下作者数量、第一作者平均下载量、第一作者机构、期刊影响因子等影响因素的作用均存在差异。相比热门分区和潜力分区，衰退分区论文年均被引频次还受到标题长度和期刊影响因子的显著正向影响。

（3）从学科总体、学科主题、主题热度分区3个层次探究年均被引频次影响因素，发现基金资助、论文篇幅、论文年龄、下载量对年均被引频次的影响均保持一致。

本研究结论对科研工作者开展科研活动、撰写论文具有一定的参考意义。①不同主题下被引频次影响因素存在差异，因此科研工作者可以根据所研究的主

题，观察当前的研究主题受哪些特定因素的影响，有针对性地进行合理规划，从而提升论文的学术影响力。②根据论文的发文趋势和引文趋势将主题归为热门、潜力、衰退分区，能够帮助科研工作者了解当前研究态势。在选择研究方向时可优先考虑热门分区主题，并对潜力分区的研究主题给予更多的关注，对于衰退分区研究主题则可通过有效的研究策略和学术合作，挖掘未来发展的新机遇。相对于热门和潜力分区，衰退分区下论文的年均被引频次还受到标题长度、期刊影响因子的影响。衰退分区主题论文应注意优化标题表达和期刊选择，从而提高学术影响力。除此之外，本研究还可以为科研评价体系提供更为多样化的评估标准，促进科研评价体系的完善。科研管理者也可以根据研究主题的特点，调整评价指标和权重，以更全面地评估研究成果的影响力，更准确地挖掘各主题下的高质量论文。

本研究仍存在一定的不足之处，如选取主题数量有限，无法全面地覆盖图书情报学领域全部的研究主题，未覆盖的研究主题可能具有不同的特征和影响因素。数据收集时间范围限定在2011—2020年，尽管覆盖了较长的研究时间，但是未能捕捉到最新的研究动态和新兴研究主题的影响。随着时间的推移，学术研究的重点可能发生变化，新兴主题可能具有不同的影响因素和发展趋势。后续研究可考虑细化引用时间，深入挖掘各主题影响因素在时间上的动态变化。此外，本研究较为全面地考虑了论文外部因素对被引频次的影响，但论文的研究方法、创新性、新颖性等内部因素未纳入分析。后续研究可利用文本挖掘和语义分析技术，将内外部因素相结合进行综合分析，并结合机器学习和深度学习模型建立预测模型，分析不同因素对预测结果的贡献。

参考文献

[1] WALTMAN L. A review of the literature on citation impact indicators[J]. Journal of Informetrics，2016，10（2）：365-391.

[2] 徐庆富，康旭东，张春博. 多期刊比较视角下的论文被引频次若干影响因素研究[J]. 情报杂志，2018，37（2）：147-153.

[3] ANTONIOU G A，ANTONIOU S A，GEORGAKARAKOS E I，et al. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature[J]. Annals of Vascular Surgery，2015，29（2）：286-292.

[4] ROTH C，WU J，LOZANO S. Assessing impact and quality from local dynamics of citation networks[J]. Journal of Informetrics，2012，6（1）：111-120.

[5] JACQUES T S，SEBIRE N J. The impact of article titles on citation hits：an analysis of general and specialist medical journals[J]. JRSM Short Reports，2010，1（1）：2.

[6] ROSSI M J，BRAND J C. Journal article titles impact their citation rates[J]. Arthroscopy：The Journal of Arthroscopic & Related Surgery，2020，36（7）：2025-2029.

[7] 张振伟，梁明修，韩锟，等. 预防医学类科技论文被引频次的影响因素分析：以《中华预防医学杂志》为例[J]. 中国科技期刊研究，2021，32（1）：125-134.

[8] LEIMU R，KORICHEVA J. What determines the citation frequency of ecological papers? [J]. Trends in Ecology & Evolution，2005，20（1）：28-32.

[9] BORSUK R M，BUDDEN A E，LEIMU R，et al. The influence of author gender，national language and number of authors on citation rate in ecology[J]. The Open Ecology Journal，2009，2（1）：25-28.

[10] 杜建，张玢，李阳. 我国医学领域不同学科作者合作度与论文影响力的关系[J]. 中华医学图书情报杂志，2012，21（3）：18-23.

[11] 王崇锋，崔运周，杨箫. 合作网络、知识网络对论文被引量的影响：基于我国管理案例研究论文的统计分析[J]. 管理案例研究与评论，2020，13（3）：356-367.

[12] 杨莉，熊泽泉，段宇锋. 基于分位数回归的期刊论文被引量预测研究[J]. 情报科学，2019，37（10）：60-66.

[13] BUELA-CASAL G，ZYCH I. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals[J]. Psicothema，2010，22（2）：270-276.

[14] FU L D，ALIFERIS C F. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature[J]. Scientometrics，2010，85（1）：257-270.

[15] YAN Y，TIAN S W，ZHANG J J. The impact of a paper's new combinations and new components on its citation[J]. Scientometrics，2020，122（2）：895-913.

[16] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[17] LDA主题抽取[EB/OL]. [2024-01-02]. https://github.com/lda-project/lda.

[18] Stata 18.0软件[EB/OL]. [2024-01-02]. https://stata.com.

[19] 胡泽文，韩雅蓉，王梦雅. 基于LDA-Word2Vec的图书情报领域机器学习研究主题演化与热点主题识别[J]. 现代情报，2024，44（4）：154-167.

[20] 陈稳，陈伟. 基于计量指标多变量LSTM模型的新兴主题热度预测研究[J]. 数据分析与知识发现，2022，6（10）：35-45.

[21] 李秀霞，程结晶，韩霞. 发文趋势与引文趋势融合的学科研究主题优先级排序：以我国情报学学科主题为例[J]. 图书情报工作，2019，63（11）：88-95.

作者简介

傅柱，男，博士，副教授，研究方向：知识组织与挖掘。
张倩，女，硕士研究生，研究方向：信息管理与信息系统。
刘鹏，男，博士，副教授，通信作者，研究方向：复杂网络分析，E-mail：liupeng19821017@126.com。

Difference of Influencing Factors of Citation in Scientific Papers from the Perspective of Subject Theme: An Empirical Study of Library and Information Science

FU Zhu ZHANG Qian LIU Peng
(School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, P. R. China)

Abstract: Examining the factors influencing citation frequency from the perspective of subject themes can provide targeted guidance for researchers in writing and optimizing their papers, and also provide a new perspective and idea for scientific research evaluation. This study uses 43 228 valid papers published in 18 CSSCI-indexed journals in library and information science from 2011 to 2020. By applying the LDA topic model to extract and identify themes from paper abstracts, and using a negative binomial regression model, we empirically investigate the factors influencing annual citation frequency from overall, thematic, and thematic classification perspectives. The findings reveal both commonalities and differences in the factors influencing annual citation frequency across subject themes. Consistent factors include funding support, paper length, paper age, and download frequency. In contrast, factors such as title length, number of keywords, and number of authors show varying impacts on annual citation frequency.

Keywords: Citation Frequency; Theme Classification; Regression Analysis; Influencing Factor; Library and Information Science

（责任编辑：王玮）