# 基于内容挖掘的学术论文影响力研究现状与展望*

辛晓梦 白如江 孔玲 王效岳
（山东理工大学信息管理学院，淄博 255049）

摘要：当前，以ChatGPT为代表的人工智能技术飞速发展，文本挖掘平台的功能优化加快有关内容挖掘的学术论文影响力研究进程，全面梳理基于内容挖掘的学术论文影响力测度的研究现状和进展迫在眉睫。通过梳理国内外利用内容挖掘方法测度学术论文影响力的研究，提出从学术、社会和技术3个维度阐述学术论文影响力的内涵。在其基础上，以时间为轴线，重点论述“学术论文影响什么、怎么影响，以及影响程度如何”的相关内容，深入阐述借助内容挖掘技术的学术论文影响力测度指标和方法。目前，基于内容挖掘的学术论文影响力测度还需利用以ChatGPT为代表的文本挖掘平台和数智技术进一步挖掘相关语义特征，深入探究引文背后的影响机制及理论，尝试从学术、社会和技术等维度，词语、句子、篇章等粒度深入文本内容，综合测度论文影响力。

关键词：学术论文；内容挖掘；学术影响力；社会影响力；技术影响力
中图分类号：G251　DOI：10.3772/j.issn.1673-2286.2024.01.003
引文格式：辛晓梦，白如江，孔玲，等. 基于内容挖掘的学术论文影响力研究现状与展望[J]. 数字图书馆论坛，2024，20（1）：23-32.

学术论文作为科研成果的重要形式，是基础研究成果和技术创新成果研究的基础。学术论文影响力作为学术评价至关重要的指标之一受到学界广泛关注。常见的学术论文影响力测度方法如文献计量法、社会网络分析法等利用外部指标测度论文的影响力，基于内容挖掘的学术论文影响力测度方法则深入文本内容挖掘学术论文的价值和贡献。人工智能、大数据等数智技术的飞速发展推动了基于内容挖掘的学术论文影响力测度研究，为科学研究良性循环与正向发展提供了技术保障。从微观上看，基于文本内容挖掘的学术论文影响力研究能够揭示高水平论文的内在规律，展现学者的科研能力和对科研活动的影响程度。从宏观上看，基于内容挖掘的学术论文影响力研究能够促进科研工作的开展，影响科研规划方向，为我国基础科学研究资助布局提供决策支持，为建立科学的评价体系、出台合理的科技政策和推动高质量的经济发展提供参考。

鉴于此，本文在系统梳理相关研究文献的基础上提出从学术影响力、社会影响力和技术影响力3个维度界定学术论文影响力的内涵。在深度调研基于内容挖掘的学术论文影响力测度研究的基础上，以时间为轴线，重点论述相关理论和指标方法、影响内容、影响方式以及影响程度，分析目前研究中存在的问题并提出建议，以期帮助学者把握研究方向、研判该研究的发展趋势，为国家制定相关政策和战略布局提供参考。

## 1 学术论文影响力的内涵

目前，影响力是学术论文评价的重要内容，学界对

收稿日期：2023-10-23
*本研究得到国家社会科学基金“基于文本内容挖掘的学术论文影响力评价研究”（编号：19BTQ085）资助。

其内涵还没有统一界定。大多数学者认为学术论文影响力是指学术论文对学术科研活动及外部社会环境所产生的影响作用，可以分为学术影响力、社会影响力和技术影响力3个维度。①学术影响力。论文的学术影响力主要是指学术论文在本学科或多学科中对科研工作者、学术活动等在方法、理论以及应用方面产生的贡献，影响范围主要为学术界。比如，van Houten等[1]认为学术影响力作为最广泛的论文影响力指标，一般是指同行科研工作者对研究者科研成果的评价和重视程度。②社会影响力。随着社会网络环境的发展，研究者逐渐将视角扩展到学术以外的社会环境，论文的社会影响力侧重于论文对文化、政策、健康、社会、大众思想和环境等方面产生的影响，影响范围主要为社会层面。Godin等[2]构建了一个包括科学、技术、经济、文化、社会、政策、环境、象征/符号与培训等因素的论文影响力模型。英国的“研究卓越框架”也将政策、服务与社会纳入论文影响力评估内容。③技术影响力。论文的技术影响力主要包括学术论文对于经济发展和技术进步的推动作用，因此其影响范围主要为经济和技术两方面。20世纪50年代，多位经济学家提出测度学术成果对生产力和经济增长的影响[3-4]。Moed等[5]提出可以利用学术影响、社会影响、技术影响、经济影响和文化影响5个方面界定学术影响力内涵。

根据以上对学术论文影响力内涵的分析，本研究把基于内容挖掘的学术论文影响力测度界定为利用人工智能、大数据等数智技术，对引文内容甚至全文本内容如词语、句子、段落、篇章、图表以及公式等进行深入挖掘，探究学术论文对科学、社会、技术等对象的影响内容、影响方式和影响程度，在此基础上测度论文的学术影响力、社会影响力和技术影响力。

## 2 基于内容挖掘的论文学术影响力研究

通过科学计量学指标、社会网络等外部指标虽然能在一定程度上反映被引论文对施引文献产生的影响，但无法展现被引论文对施引文献的真正影响内容。随着数智环境和全文本计量时代的到来，包含不同粒度数据源信息的结构化全文数据日益丰富，更能满足对细粒度客体和指标语义特征等的更精细化的需求，基于内容挖掘的论文学术影响力研究成为学术论文影响力研究的主力军。论文学术影响力的产生过程和内容如图1所示。基于内容挖掘的论文学术影响力研究在一定的理论支撑下，除了分析作者、期刊声望、合作关系等外部因素的影响，同时能够挖掘出被引论文对施引文献在新方法、新理论、新技术、新主题等创新性内容上的影响，因此论文的创新性测度也是论文学术影响力测度的重要部分。具体来说，基于内容挖掘的论文学术影响力研究包括测度理论、测度方法和论文创新性测度。

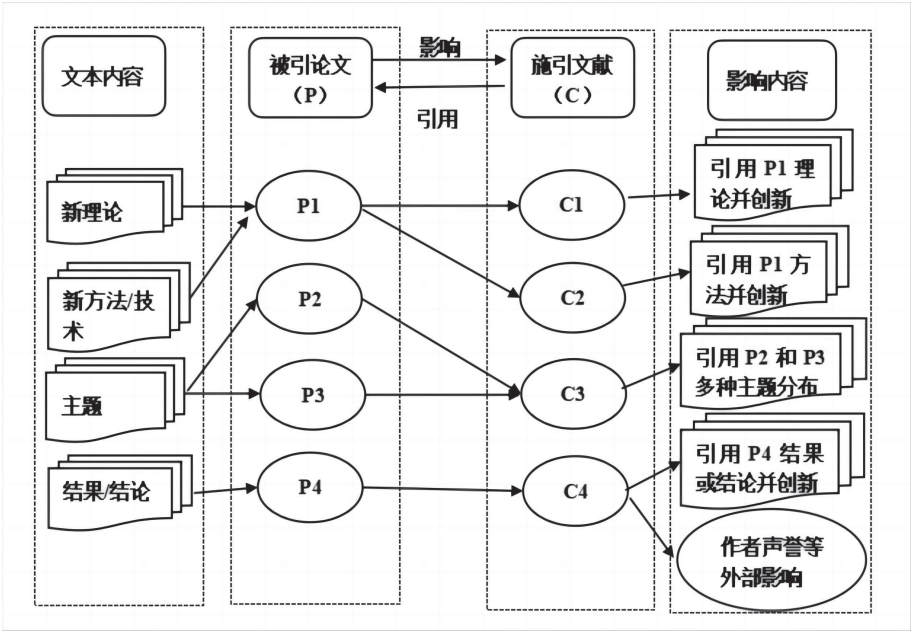


图1 论文学术影响力的产生过程和内容

## 2.1 测度理论

在以往研究中，学者们已广泛将生物学、语言学、计算机学以及社会学等多学科理论应用到论文学术影响力研究中。未来，基于全文计量分析的学术论文影响力研究越来越需要其他学科的知识来不断丰富自身的理论和知识架构。

2017—2022年，学者们根据学科理论的自身逻辑特点将其应用于论文学术影响力测度指标或模型构建。有些研究基于理论与论文影响力分析逻辑的相似性，直接移植理论公式来进行学术论文影响力测度，比如：徐浩[6]应用意义建构理论和知识谱系理论，分别从行为角色角度和知识扩散角度分析学术论文中知识的跨学科流动情况，以此测度论文学术影响力；梁国强等[7]借用创新扩散理论，从知识扩散和流动角度将学科多样性作为论文影响力宽度的测度指标，并直接应用于论文影响力分析。以上理论大多来源于信息情报、新闻等学科，其概念和指标具有通用性，在分析学科之间的知识流动和传播方面具有一定的优势。后续可以考虑将生物学、计算机等学科的理论合理应用到论文学术影响力分析中。

另外，有些学者在所借鉴理论的基础上进行指标、公式等的创新，以此测度论文的学术贡献和价值。牌艳欣等[8]基于生态系统和学术交流系统的相似性，将学术论文影响力分为原生影响力和次生影响力，根据生态位理论中的生态位宽度公式构建基于论文内容创新性、线上关注度、线下关注度3个维度的学术论文生态位宽度测度模型。楼雯等[9]提出，安娜·卡列尼娜原理的系统性思想与从全面、多样的角度测度学术论文影响力的逻辑相通，基于此从内容价值、学术价值、应用价值和社会价值4个维度构建评价指标，测度学术论文影响力。认知计算理论能够利用计算机模拟人脑，处理和识别非结构化数据，并进行信息无偏见评价，或将成为未来学术论文影响力测度的新角度。

以上研究分别从数据处理、论文价值挖掘、论文贡献测度以及评价全面性等角度，将不同学科的理论体系应用在论文学术影响力评价中，进一步提高了对学术论文语义层面的关注度，丰富了学术论文影响力评价体系。不过，当前应用其他学科理论测度论文学术影响力的研究还比较少，需要引入更多成熟的理论方法，取各学科之精华并借鉴其研究特长和研究路径，不断深化研究。

## 2.2 测度方法

目前基于内容挖掘的论文学术影响力测度方法主要包括基于规则和人工识别的方法、基于传统机器学习的方法，以及基于深度学习的方法。

（1）基于规则和人工识别的方法。应用于学术论文影响力测度的规则方法主要包括词频统计和分布统计。词频能够体现词语在文本中的重要程度，最为直接和简单，但仅依靠引用内容中的短语测度论文影响力的方法存在无法实现定量分析的问题[10]。针对此问题，可以考虑利用向量空间模型、信息熵、引用功能赋予权重，以及结合上述方法综合评价论文质量的办法。在指定抽取规则抽取研究评价句的基础上，从学术评价句类型和位置角度分析论文影响力的方法虽然能够更全面地表达文本内容，但存在数据和研究领域范围较小以及抽取正确率需要进一步提高的问题[11]。有研究通过人工识别的方式试图解决识别结果不准确的问题，从全文本角度选取被引对象出现频次、当年影响因子和下载量作为论文学术影响力测度指标，并构建单篇论文的学术影响力测度模型[12]，该研究得出不同领域论文的差异性较大的结论。随后，谢珍等[13]结合引用情感、强度、主题相似度、引用位置、引用主体重要性等6个维度的引用特征组合，区分被引论文对施引文献的影响，该方法能更细致地剖析论文的被引特征，使论文价值计算更为精确和全面。

词语、句子、全文本等不同粒度的内容能够表达不同角度的论文影响力测度理念，同时，引用强度、引用位置、引用情感以及主题相似度等不同的引用特征也能够展现论文对施引文献产生的不同影响。规则抽取方法较为简单和直接，但容易出现识别结果不准确的问题。人工识别方法虽然准确率较高，但抽取效率还需要进一步提升。因此，在对论文的内容进行特征分析时，应注意考虑不同语法、语义甚至语境特征，借助更多人工智能技术解决文本抽取和分析方面的问题，以更全面、深入地分析论文学术影响力。

（2）基于传统机器学习的方法。与基于规则的方法相比，传统机器学习作为一种使用计算机构建概率统计模型的方法，能够充分挖掘、预测和分析论文内容数据，更能适应基于不同数据集的论文学术影响力测度，且更加规范。当前研究主要通过关键词挖掘、语言模型和主题模型进行分析。

首先，耿树青等[14]使用SVM算法对引用内容进行

情感分类，并提出一种基于“被引次数-引用情感”指标的引用情感测度方法，为论文的学术影响力测度提供了新思路。该方法虽然与传统方法相比具有优势，但同时也存在引用内容抽取困难、缺少提供引用内容的语料库以及自动情感分类困难等问题。其次，语言模型中的Word2Vec模型也可进行文本挖掘和知识发现。可通过Word2Vec模型对文献进行建模、表征，并基于词向量技术从文献中挖掘潜在知识，从而发现论文中具有潜在价值的新观点、新技术[15]。再次，基于朴素贝叶斯方法[16]的稳定性和算法简单的优势，构建研究问题、理论、方法和结论4个知识元本体的理论与方法分类模型，并将其作为论文影响力测度的4个维度。该方法虽然具有一定可行性，但需要进一步调整相关算法和抽取规则。姜霖等[17]通过情感引用中的语义单元进行情感量化，用情感值来衡量单篇文献的学术影响力。最后，论文主题是学术论文内容高度凝练的结果，是体现论文学术影响力的重要特征之一。主题模型解决了文本内容分类问题，能够挖掘文本深层次信息，在应用到论文影响力研究中时能有效弥补标题、摘要和关键词的不足，更能代表被引文献的研究内容。通过有监督的主题建模方法发现主题并对创新指数排序，能有效测度论文的学术影响力[18]。

传统机器学习算法种类多样，能够较为有效地应用于论文影响力测度。可以尝试引入更多的传统机器学习方法，例如基于Node2Vec模型分析学术论文群落影响力测度的规范化指标等。

（3）基于深度学习的方法。近几年，随着深度学习和神经网络相关技术的不断成熟，深度学习方法作为一种无监督特征学习和特征层次结构学习方法，在文本抽取和识别效果方面具有优越性，部分学者引入相关的算法进行论文学术影响力测度研究。①采用规则匹配和BERT模型相结合的方法抽取能够表征论文价值的学术创新贡献句。这种方法虽然能够有效抽取出贡献句，但还存在抽取结果不准确的问题[19]。②BIOBERT模型。相较于BERT模型，该模型对正负样本的识别能力更高，鲁棒性更高，因此通过利用此模型对引用语句进行词频统计并结合人工筛选获取表征突破性评价的常用词，能够有效识别突破性文献，但其作为生物医学领域的特异性模型也存在其他领域不适用的问题[20]。③利用ALBERT方法从全文本中抽取意义完整的贡献句，通过贡献词语揭示论文贡献，从而实现语义化、智能化的学术论文影响力和创新力测度[21]。该方法主要通过抽取贡献词和贡献句分析学术论文的贡献，展现学术论文的价值。上述研究大多采用深度学习中的某种方法或方法组合，通过抽取和分析论文文本中的贡献短语或贡献句，探索论文影响力的测度方法或模型，并取得了较为良好的效果。

基于内容挖掘的论文学术影响力测度方法不但能够反映论文之间的引用关系，还能够准确揭示学术论文被引用了哪些内容、为什么被引用、如何影响施引论文以及影响程度。因此，随着文本挖掘技术、人工智能、深度学习等技术的不断发展，后续应进一步探究学术论文影响了什么、怎么影响以及影响程度如何的问题。在影响内容方面，揭示被引论文对施引论文理论、方法、结论等的作用；在影响程度方面，施引论文的发文机构、学科分布、基金支持等能够体现被引论文对施引论文的具体影响力；在影响方式方面，一次引用、三角引用以及多代引用关系展现了被引论文对多篇论文的影响路径。

## 2.3 创新性测度

学术论文提出的新方法、新理论、新技术得到施引论文的引用，这在一定程度上体现了学术论文对施引论文的影响方式和影响程度。因此，学术论文创新性是论文学术影响力的一个表现特征。目前，学者们主要从学术论文的主题和创新贡献句角度来探究学术论文创新性的测度方法和指标。

学术论文的主题是其内容的高度凝练，通过测度主题的新颖性来判断学术论文的创新性的方法具有天然的优势，基于此，学者们从主题分析入手构建学术论文创新性的测度指标和方法。许丹等[22]基于对学术论文主题的分析，构建文档主题新颖度指标，通过自然语言方法，对于词语出现的频率和趋势规律进行统计运算，分析出该篇论文在整个论文集内所体现的新颖程度。Mairesse等[23]基于时间的角度，认为学术论文中包含创新性知识片段，提出利用该创新性知识片段出现的年份测度学术论文的创新性。逯万辉等[24]基于Doc2Vec算法和隐马尔可夫链构建主题新颖度测度模型以测度学术论文的创新性。杨京等[25]对论文的研究主题和前沿主题进行相似度对比计算，提出了一种基于研究主题对比的学术论文创新性测度方法，结果表明该方法能够有效、迅速、准确地从论文内容各角度对学术论文的创新性进行测度。此外，结合时间维度和主题

分析构建的创新指数能够在一定程度上反映论文的创新性和影响力[26]。

随着研究的深入，学者们逐渐将研究目光扩展到主题以外的句子层面，从学术论文的摘要甚至全文本内容角度探究学术论文创新性的测度方法和指标。章成志等[11]通过选取标志词、指定抽取规则抽取创新研究评价句的方法，将创新研究的内容分为观点发现、概念理论、模型方法、派别领域、系统软件以及实践应用6类。周海晨等[19]提出基于标注创新贡献短语的学术创新贡献识别方法，制定相关细粒度的抽取规则，并成功运用到大规模数据集中。基于神经网络探测学术论文内容的创新性的方法表现较为突出[27]，但目前还未在论文学术影响力测度中得到广泛的应用。另外，Chen等[28]提出基于N-grams模型从论文摘要中自动提取创新点的方法来探测学术论文的创新思想，该方法能够通过适当组合标识符提高对创新点的自动提取性能。通过Word2Vec模型挖掘学术论文中的潜在知识，能够有效发现论文中的具有潜在价值的新观点、新技术[15]。罗卓然等[21]利用ALBERT方法从全文本中抽取意义完整的贡献句，通过贡献词语揭示论文贡献，实现语义化、智能化的论文学术影响力和创新性测度。

## 3　基于内容挖掘的论文社会影响力研究

近年来，随着开放科学运动的兴起和发展、社交网络的广泛应用，学术成果的受众群体更加广泛。社会各界不但想了解学术论文在学术界体现的学术价值，而且也关注学术论文的成果是否对解决社会问题、促进文化繁荣以及优化经济环境等产生积极的作用并带来社会效益。普通公众对科学研究活动的参与度和关注度提升，推动科研管理部门将学术论文影响力的测度视角从学术影响力层面拓宽到社会影响力层面，学术论文成果的社会影响逐渐成为吸纳公众资金和研究支持的重要因素。社交媒体类数据源和工具是基于内容挖掘的论文社会影响力测度的研究基础，因此充分挖掘和拓展方便、可用的数据来源和工具，成为现有研究需要重点关注的问题。同时，社交媒体计量时代的到来使得学术论文与非传统学术环境中的社会活动和大众思想等产生了联系，并通过公众的思想变化和行为改变等对社会环境产生了不同程度的影响，这些影响内容和方式成为基于内容挖掘的学术论文社会影响力测度的重要内容。因此，从数据源或工具、影响内容、影响方式等角度对论文的社会影响力研究进行梳理。

### 3.1　数据源和工具

Altmetrics在2010年由美国学者Priem等[29]提出，是利用社交媒体指标测度学术论文社会影响力的信息计量学分支。Altmetrics作为一种基于社交网络、新闻媒体和在线文档管理软件的综合文档使用跟踪和评估方法，可以量化论文的社会传播和使用情况。目前，国内外用于论文社会影响力分析的Altmetrics数据源主要包括原始数据源和数据整合分析工具两类。常用的原始数据源包括社交媒体、新闻媒体、在线学术网络、学术社区等网站和文献管理、百科等平台。数据整合分析工具则主要包括Altmetric.com、PLoSALM、PlumX、Kudos和Impact Story等。学者们利用这些数据源或工具收集网络文本数据加以分析，大多从主流媒体、研究学者以及普通公众的角度，通过挖掘用户对被引论文的报道、分享、引用和讨论等行为，反映社会各界对学术论文的认可和关注程度。例如，学术论文通过新闻媒体的报道而影响社会大众思想和观念，通过政策性文件提及将其理论或方法应用于政策制定，以及通过Facebook、Twitter等社交媒体引发社会公众对成果的关注等。总的来说，学术论文基于网络媒体平台，通过不同的方式影响不同领域。

首先，虽然许多研究人员通过构建Altmetrics指标，取得了良好的测度效果，但大多数研究对指标的选取和定义还比较简单和粗放。其次，Altmetrics指标大多涉及提及、分享等方面，复合性指标在实践中的应用不足，未来研究还需深入语义层面深度分析学术论文的社会影响力。最后，除Altmetrics指标外，当前研究中用于学术论文社会影响力测度的其他维度指标还比较少，如文本内容中的情感类、观点类指标，这可能成为未来研究的新方向。

### 3.2　影响内容和方式

论文的社会影响力主要体现在学术界以外的社会环境中，影响范围涉及社会、文化、健康、思想以及政策等多个方面。多个国家针对论文的社会影响力开展了相关研究，其中具有代表性的主要有英国、澳大利亚、美国以及荷兰等。这些国家的相关研究着重考虑论文对

社会、文化、政策、安全等领域产生的影响作用。

除了通过制定评估框架的方式体现论文影响力的影响范围以外，近年来，案例研究成为国外机构评估学术论文成果社会价值的有效方法，评估策略通常包括分析影响的途径和领域、特定学科影响两个方面。在分析影响途径和领域方面，通过主题分析能够探究学术论文的影响途径、受益人以及对不同学科领域的影响作用。总的来说，一般将论文的社会影响力传播途径分为研究吸收、研究使用和研究影响力3类[30]，在此基础上有学者提出将其分为传播、获取和使用3类[31]，而赵蓉英等[32]提出将其划分成更详细的使用型、捕获型、提及型、社交媒介型和引文型。影响领域涉及政策、教育等多个领域。在特定学科论文的社会影响研究方面，学者们主要通过案例文本内容分析和主题建模等方法对护理学[33]、健康科学[34]以及人文科学[35]等学科进行影响内容或领域分析。虽然案例研究的结果不具有普遍性，但能深度揭示学术论文的社会影响力，因此案例研究能够更好地厘清社会影响力的产生过程和路径。国外的一些案例研究已经采用文本内容挖掘技术，但对影响的具体性质的关注仍然不足。国内部分学者开始尝试进行案例研究，但还面临不同学科指标通用性以及长期适用性不足等挑战。未来应进一步尝试提高研究结果的代表性，保持与时俱进状态，并考虑对指标赋予权重，增强可操作性。

对于影响内容而言，社会影响力展现了被引论文对大众思想、政策、健康以及社会文化等的影响作用；对于影响程度而言，社交媒体软件中公众的下载、转发和分享的频次，以及评论观点和情感等指标可以展现论文在社会层面的影响广度和强度；对于影响方式而言，从公众接触和参与到意识变化，再到行为改变，展现了论文影响由浅入深的3个层次。

## 4　基于内容挖掘的论文技术影响力研究

当前，科学研究与技术创新之间的联系日益紧密，越来越多专利直接引用科学论文从而催生出更多的技术成果，更有许多科学家在发表学术论文的同时将科研成果申请为专利。专利作者出于不同的目的和动机引用学术论文。通过分析学术论文向专利的知识流动情况如专利的主题分布，学术论文所影响的专利价值实现情况如经济价值、法律状态以及新技术产生情况等，能够充分反映学术论文对专利的影响内容和程度。

对专利引用论文的文本内容进行分析，能够表示科学与技术在内容上的联系，从而识别出学术论文在语义层面的隐性知识，得到论文对专利真正具有影响力的文本内容。由于专利文本的特殊性，论文对专利的影响程度需要用特定的方法指标来度量。1994年，Narin[36]提出将文献计量学方法应用于论文对专利的影响研究领域，Meyer[37]进一步对该方法进行研究。目前，学者们通过利用模型、引用特征、主题分析以及内容相似度等对论文的技术影响力进行分析。

（1）基于模型的分析。即在一定的理论支撑下，利用相关模型分析论文的技术影响力。例如，裴云龙等[38]基于专利对科学文献的引用和重组搜索理论，构建了基于学术论文成果直接和间接影响专利有效性的概念模型，研究了中外科学文献对中国高新产业技术创新质量的影响。卞雅莉[39]则利用负二项回归模型，通过引入专利引用论文的质量来探索纳米领域科学对技术的影响作用。由于模型具有深厚的理论基础，基于模型探讨论文对专利和技术的影响作用的方法具有优越性，后续可引入因果模型等相关理论模型进一步探讨学术论文技术影响力的产生路径和技术发展的因果关联。

（2）基于引用特征的分析。多位学者从引用动机、引用主体、引用目的、引用功能等方面对专利引用进行了分析。专利的引用功能和动机通常可解释技术的传承性、其他技术的缺陷，以及验证引用的真实性。Huang等[40]为了确保专利的创新性，提出在一定前提条件下可把专利引用文献抽象为知识流动的形式。Li等[41]从不同引用主体的引用动机角度来区别引用的真实性，以确定被引论文对施引专利的正负向影响。赵阳等[42]提出利用引证动机反映专利引用论文的真实原因，通过分析引用是否真实、是否正确以及两者内容是否相关的客观信息来确定被引论文对施引专利的影响。在学术论文影响力测度的实际应用中，通过引用特征能够有效考察施引作者的引用动机和目的，能够更为精准地评估学术论文对施引专利所产生的影响力。

（3）基于主题的分析。主题映射能够反映科学与技术间的演化关系，预测有潜力转化为商业应用的论文研究重点主题。张金柱等[43]将专利科学引文的关键词和学科分类作为专利引用科学知识的表征进行特定技术领域内的突破性创新识别，在基因工程领域验证了该方法的有效性。有学者进一步通过识别被引论文的科学知识主题遴选突破性创新，在多个学科交叉领

域获得了良好的实验结果，但科学知识抽取和匹配准确率需要进一步提升[44]。丁文晴等[45]利用Python语言，通过LDA主题模型进行文本挖掘，识别科技演化模式并与市场主体对接，结果证明该方法具备稳健性。学术论文的主题是学术论文内容的高度凝练，通过对学术论文和施引专利的主题分析，能够展现学术论文对施引专利产生的影响内容是什么以及这种影响是如何产生的，从而分析学术论文对专利研究方向、技术进步和经济发展等的影响作用。

（4）基于内容相似度的分析。通过摘要等文本内容的相似度计算，对比被引论文和施引文献的内容关联度，从而明确论文对专利的影响路径和技术演进脉络。胡一鸣[46]通过挖掘专利科学引文内容层面的深层信息，将专利科学引文的内容挖掘结果与专利的内容挖掘结果进行对比，最终实现内容层面的科学-技术关联关系分析。张雪等[47]从科技关联度、主题、循环周期等角度分析专利对论文的引用情况，研究结果为提高领域的基础研究和技术创新能力提供了有效建议。冯立杰等[48]利用引文网络和语义分析方法识别技术演化路径，并将被引论文和施引专利通过语义相似性计算连接到技术演化路径的末端，揭示了论文对技术演进趋势的影响。基于内容相似度的相关研究主要从学术论文和施引专利间的知识流动情况和技术演进情况等角度进行分析，能够更为深入地探讨论文对技术的影响路径。

以上研究大多采用主题和引用内容方法，通过探索论文与专利的作用关系和路径，有效识别出论文对科技创新的影响。但从文本内容层面分析论文对专利影响作用的研究主要集中在语义相似度计算和主题分析两个方面，相对来说更加注重专利的现时效益。但由专利转化的技术应用成果的效益也能够反映学术论文的影响力，经济效益等也是测度学术论文对技术创新影响程度的有效指标。后续考虑从论文的现时影响和后续影响两个方面综合测度其技术影响力。

## 5　基于内容挖掘的学术论文多维综合影响力研究

针对不同领域、不同对象，学术论文具有学术、社会和技术等多维影响力，除单一维度的测度研究以外，有学者融合学术和社会两个维度对论文影响力进行了综合测度。

部分研究单纯将学术影响力和社会影响力指标进行融合。例如，彭秋茹等[49]利用引文指标与Altmetrics指标，对报纸论文影响力进行多维测度。Hou等[50]运用层次分析法将学术影响力指标与社会影响力指标结合，计算学术论文综合影响力，比较高影响力论文与高被引论文的特征，并进一步探究论文特征对社交媒体指标的影响。张靖雯等[51]以引文扩散为视角，试图从论文学术影响力与社会影响力两个方面综合探究不同引文扩散模式下论文的影响力情况，为学术论文影响力测度研究构建多维度指标、提升学术影响力提供了借鉴参考。此外，也有多位学者加入其他指标共同构建测度模型。例如，许鑫等[52]从潜在影响力、学术影响力、社会影响力3个维度构建多指标融合体系，对数据论文影响力进行综合测度。杨思洛等[53]在传统二维测度模型基础上融入全文本分析，通过层次分析和主成分分析，将引文和Altmetrics对应的次数类、位置类和情感类全文本指标融合为学术和社会影响力指标。邱均平等[54]从科学交流视角出发，从原生影响力、Web 1.0影响力和Web 2.0影响力3个方面分析学术论文影响力的形成机理，其构建的测度模型具备创新性与可行性。

基于文本内容的论文影响力多指标综合测度研究主要融合学术影响力和社会影响力，其应用的学术论文影响力测度指标大多为引用特征中的语法特征指标，相关研究甚至从全文本的角度对学术和社会影响作用进行深入剖析。然而，学术论文不仅对学术界和社会环境产生影响，还会对经济发展和技术进步等方面产生影响。因此，上述研究虽然在一定程度上能够拓展测度视角，但融合方法较为单一，且与技术影响力的融合程度不高。后续可考虑融入技术影响力，从全文本角度测度3个维度的综合影响力。

## 6　基于内容挖掘的学术论文影响力研究的不足与展望

综上可知，学术界在基于内容挖掘的学术论文影响力测度研究方面已经取得丰富的研究成果，但随着社会发展和技术进步，基于内容挖掘的学术论文影响力研究的广度和深度还有待进一步拓展。

（1）内容贡献的语义化特征需要进一步挖掘，以深入揭示学术论文影响了什么。第一，全文计量研究还处于从关注语法特征到关注语义特征的过渡阶段。目前研究主要聚焦文中提及频次、引用位置等语法层面，

对引用情感、行为动机等语义层面的分析相对较少。第二，语法特征与语义特征分析大多相互独立，比如，引用强度只能展现文献重要性，无法揭示每次引用时作者的目的和动机，导致相关研究忽视了学术论文内在的语义逻辑，无法充分挖掘论文所影响的新理论、新方法、新思想，从而无法构建完整的评价体系。因此，应进一步挖掘论文影响力指标的语义和语用特征。一方面，重视语义和语用特征的应用，发挥数智赋能效用，充分利用以ChatGPT为代表的文本挖掘平台的相关功能和服务，挖掘引用情感、引用动机以及引用功能等语义特征，改良传统的评价体系；另一方面，提取引用位置等语法特征和引用动机、情感等语义特征的共性，充分结合两者进行分析，识别学术论文对施引文献的真正影响内容。

（2）引用背后的影响机制需要进一步探讨，理论支撑需要进一步拓展。首先，图书情报领域理论基础不够牢固，主要体现在概念类型不统一和测度理论较少等问题上。其次，利用其他学科理论进行论文影响力测度的研究有局限性，主要关注语言学、社会学及生物学等领域的理论。最后，只注重研究结果的展示，忽略了具体影响过程和途径。目前的论文影响力测度研究只通过分析数据来解释论文之间的关系、反映数据的相关规律，无法体现论文引用背后的因果关系。因此，深入探究引文背后的影响机制及理论成为当前工作的重点。第一，立足于新时代的科学成果评价，加快形成对论文影响力概念类型的共识。第二，鉴于已有研究达到了良好效果，应进一步挖掘应用多种学科理论和方法，比如：物理领域的路程计算公式，经济学领域的马太效应等相关理论，语言学领域的因果推论、论证挖掘等相关理论。尤其是通过因果推论，能够发现学术文本中学科间的知识流动和因果关联关系，揭示科学知识的演化路径，探索背后的影响机制和演化规律，理解引用行为的内在机理，进而分析学术论文为什么被引用、如何被引用等问题。

（3）测度影响程度的指标和融入技术影响力的综合影响力指标需要进一步丰富。一方面，现有研究多关注论文的学术影响力，对于社会影响力和技术影响力的关注度还不够，这将导致学术论文影响力研究通常固定在学术界，很难扩展到社会层面和成果转化层面。另一方面，在利用多维度论文影响力构建测度模型时，由于专利文本的特殊性，研究者多基于学术影响力和社会影响力二维影响力构建测度指标，技术影响力的参与度较低，且技术影响力方面的影响程度测度指标研究还不够深入。对此，应尝试从论文学术影响力、社会影响力、技术影响力等多个维度，词语、句子、篇章等多种粒度综合测度论文影响力，提升有效测度能力。首先，借助生成式模型等先进的计算机技术和国内外数据库与挖掘平台的相关功能，识别和分析词语、句子、篇章等多种粒度内容。其次，在学术影响力方面，综合考虑发文特征如机构、作者、时间、学科分布以及基金支持等，分析具体的论文学术影响力；在社会影响力方面，从情感类、次数类、观点类、主题类等学术论文内容质量和学术价值指标方面深层次探讨社会影响力的真正内涵；在技术影响力方面，基于分支主题、专利经济价值、专利类型等，识别论文对专利的实质影响力，并尝试多种方法对其赋予权重。最后，基于3个维度从全文本分析的角度构建语法特征、语义特征甚至语用特征相结合的论文影响力综合测度模型，使学术论文影响力的测度指标更综合和全面、体系更合理。

## 参考文献


[1] VAN HOUTEN B A，PHELPS J，BARNES M，et al. Evaluating scientific impact[J]. Environmental Health Perspectives，2000，108（9）：A392-A393.

[2] GODIN B，DORÉ C. Measuring the impacts of science：beyond the economic dimension[R/OL]. Helsinki：Helsinki Institute for Science and Technology Studies，2005[2023-10-19]. http://www.csiic.ca/PDF/Godin_Dore_Impacts.pdf.

[3] MANSFIELD E. Academic research and industrial innovation[J]. Research Policy，1991，20（1）：1-12.

[4] NARIN F，HAMILTON K S，OLIVASTRO D. The increasing linkage between U.S. technology and public science[J]. Research Policy，1997，26（3）：317-330.

[5] MOED H F，HALEVI G. Multidimensional assessment of scholarly research impact[J]. Journal of the Association for Information Science and Technology，2015，66（10）：1988-2002.

[6] 徐浩. 融合研究工具及方法的跨学科知识扩散过程及测度[D]. 南京：南京大学，2017.

[7] 梁国强，侯海燕，高桐，等. 基于创新扩散理论的学术论文影响力广度研究[J]. 图书情报工作，2019，63（2）：91-98.

[8] 牌艳欣，周彦廷. 生态位理论视角下学术论文影响力评价研究[J]. 情报理论与实践，2022，45（12）：119-127，145.

[9] 楼雯，刘小曼，蔡蓁. 基于安娜·卡列尼娜原理的单篇科学论文评价方法研究[J]. 情报理论与实践，2022，45（7）：89-94，41.

[10] 刘盛博，王博，唐德龙，等. 基于引用内容的论文影响力研究：以诺贝尔奖获得者论文为例[J]. 图书情报工作，2015，59（24）：109-114.

[11] 章成志，李铮. 基于学术论文全文的创新研究评价句抽取研究[J]. 数据分析与知识发现，2019，3（10）：12-19.

[12] 翟姗姗，叶丁菱，胡畔，等. 融合Altmetrics与引文分析的数据论文学术影响力评价[J]. 情报学报，2020，39（7）：710-718.

[13] 谢珍，马建霞，胡文静. 基于多维度引用特征的学术论文评价研究[J]. 情报理论与实践，2023，46（8）：51-58，76.

[14] 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.

[15] TSHITOYAN V，DAGDELEN J，WESTON L，et al. Unsupervised word embeddings capture latent knowledge from materials science literature[J]. Nature，2019，571：95-98.

[16] 李贺，杜杏叶. 基于知识元的学术论文内容创新性智能化评价研究[J]. 图书情报工作，2020，64（1）：93-104.

[17] 姜霖，张麒麟. 基于引文细粒度情感量化的学术评价研究[J]. 数据分析与知识发现，2020，4（6）：129-138.

[18] LIU X Z，ZHANG J S，GUO C. Full-text citation analysis：a new method to enhance scholarly networks[J]. Journal of the American Society for Information Science and Technology，2013，64（9）：1852-1863.

[19] 周海晨，郑德俊，郦天宇. 学术全文本的学术创新贡献识别探索[J]. 情报学报，2020，39（8）：845-851.

[20] 王雪，杨雪梅，林紫洛，等. 基于引文全文本的医学领域突破性文献识别研究[J]. 情报杂志，2021，40（3）：132-138.

[21] 罗卓然，蔡乐，钱佳佳，等. 学术论文创新贡献句识别研究[J]. 图书情报工作，2021，65（12）：93-100.

[22] 许丹，徐爽，陈斯斯，等. 基于自然语言词对法的文献主题新颖性探测研究[J]. 图书情报工作，2018，62（8）：130-138.

[23] MAIRESSE J，PEZZONI M. Novelty in science：the impact of French physicists' novel articles[C]//23rd International Conference on Science and Technology Indicators. 2018：212-220.

[24] 逯万辉，苏金燕，余倩. 学术成果主题新颖性与学术引用的相关关系研究[J]. 情报资料工作，2018（6）：68-73.

[25] 杨京，王芳，白如江. 一种基于研究主题对比的单篇学术论文创新力评价方法[J]. 图书情报工作，2018，62（17）：75-83.

[26] 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.

[27] AMPLAYO R K，HONG S，SONG M. Network-based approach to detect novelty of scholarly literature[J]. Information Sciences，2018，422：542-557.

[28] 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.

[29] PRIEM J，HEMMINGER B H. Scientometrics 2.0：new metrics of scholarly impact on the social web[J]. First Monday，2010，15（7）：2.

[30] CATHERINE M S. Exploring and assessing social research impact：a case study of a research partnership' s impacts on policy and practice[D]. Edinburgh：University of Edinburgh，2012.

[31] 余厚强，邱均平. 替代计量指标分层与聚合的理论研究[J]. 图书馆杂志，2014，33（10）：13-19.

[32] 赵蓉英，吴胜男，王旭，等. Altmetrics理论与实践[M]. 北京：科学出版社，2019.

[33] KELLY D，KENT B，MCMAHON A，et al. Impact case studies submitted to REF 2014：the hidden impact of nursing research[J]. Journal of Research in Nursing，2016，21（4）：256-268.

[34] GREENHALGH T，FAHY N. Research impact in the community-based health sciences：an analysis of 162 case studies from the 2014 UK Research Excellence Framework[J]. BMC Medicine，2015，13：232.

[35] 曾粤亮，郑汉，杨思洛. 人文科学研究成果的社会影响探析：以REF案例为样本[J]. 情报学报，2023，42（7）：842-856.

[36] NARIN F. Patent bibliometrics[J]. Scientometrics，1994，30（1）：147-155.

[37] MEYER M. What is special about patent citations? differences between scientific and patent citations[J]. Scientometrics，2000，49（1）：93-123.

[38] 裴云龙，蔡虹，赵皎卉. 纳米科学对纳米技术的影响：基于NPR的分析[J]. 情报杂志，2010，29（10）：1-4.

[39] 卞雅莉. 科学引文对企业专利质量的影响：以纳米材料产业为例[J]. 情报杂志，2013，32（1）：50-54.

[40] HUANG C，NOTTEN A，RASTERS N. Nanoscience and technology publications and patents：a review of social science studies and search strategies[J]. The Journal of Technology Transfer，2011，36（2）：145-172.

[41] LI R，CHAMBERS T，DING Y，et al. Patent citation analysis：calculating science linkage based on citing motivation[J]. Journal of the Association for Information Science and Technolo-

gy，2014，65（5）：1007-1017.

[42] 赵阳，文庭孝. 专利引证动机分析[J]. 情报理论与实践，2017，40（7）：28-32，16.

[43] 张金柱，张晓林. 基于被引科学知识主题突变的突破性创新识别[J]. 现代图书情报技术，2016（S1）：42-50.

[44] HEFFERNAN K，TEUFEL S. Identifying problems and solutions in scientific text[J]. Scientometrics，2018，116（2）：1367-1382.

[45] 丁文晴，崔晶，马俊红，等. 科技演化模式与市场的距离测度：以我国生物制药领域为例[J]. 情报理论与实践，2020，43（6）：142-148.

[46] 胡一鸣. 基于表示学习的专利科学引文元数据自动抽取及其内容挖掘研究[D]. 南京：南京理工大学，2018.

[47] 张雪，张志强，陈秀娟，等. 合成生物学领域的基础研究与技术创新关联分析[J]. 情报学报，2020，39（3）：231-242.

[48] 冯立杰，周炜，刘鹏，等. 基于引文网络和语义分析的技术演化路径识别及拓展研究[J]. 情报理论与实践，2023，46（1）：90-99，131.

[49] 彭秋茹，阎素兰，杨波，等. 融合引文与Altmetrics的报纸论文影响力综合评价方法研究[J]. 图书与情报，2018（5）：11-21，28.

[50] HOU J H，MA D. How the high-impact papers formed? a study using data from social media and citation[J]. Scientometrics，2020，125（3）：2597-2615.

[51] 张靖雯，闵超. 引文扩散视角下论文学术影响力和社会影响力比较研究：以生物医学为例[J]. 情报学报，2023，42（1）：31-42.

[52] 许鑫，叶丁菱. 多维影响力融合视域下的数据论文评价研究[J]. 情报学报，2022，41（3）：275-286.

[53] 杨思洛，聂颖. 结合全文本分析的论文影响力评价模型研究[J]. 现代情报，2022，42（3）：133-146.

[54] 邱均平，刘亚飞，魏开洋. 科学交流视角下学术论文影响力多维评价[J]. 情报理论与实践，2023，46（6）：47-54.

## 作者简介

辛晓梦，女，硕士研究生，研究方向：文本挖掘与科学计量评价。
白如江，男，博士，教授，通信作者，研究方向：文本挖掘与科技情报分析，E-mail：brj@sdut.edu.cn。
孔玲，女，博士，讲师，研究方向：科技情报分析、文本挖掘与信息计量。
王效岳，男，博士，教授，研究方向：科技情报分析与文献挖掘。

Current Status and Prospect of Research on the Influence of Academic Papers Based on Content Mining

XIN XiaoMeng BAI RuJiang KONG Ling WANG XiaoYue
(School of Information Management, Shandong University of Technology, Zibo 255049, P. R. China)

Abstract: At present, with the rapid development of artificial intelligence technology represented by ChatGPT, the functional optimization of the text mining platform has accelerated the research process of academic paper impact based on content mining, and it is urgent to comprehensively sort out the research status and progress of academic paper impact measurement based on content mining. By combing the research on the influence of academic papers based on content mining at home and abroad, this paper proposes to explain the connotation of the influence of academic papers from three dimensions: academic, social, and technological. On this basis, this paper focuses on the relevant content of "what academic papers affect, how it affects, and how much it influences" based on the time axis, and expounds the indicators and methods of measuring the influence of academic papers based on content mining. At present, the impact measurement of academic papers based on content mining needs to use the text mining platform represented by ChatGPT and digital intelligence technology to further explore the relevant semantic features, deeply explore the influence mechanism and theory behind citations, and try to comprehensively measure the influence of papers from multiple dimensions such as academic, social, and technological, and the granularity of words, sentences, and articles.

Keywords: Academic Paper; Content Mining; Academic Influence; Social Influence; Technological Influence

（责任编辑：王玮）