

## 论文新颖性测度研究综述

柴嘉琪<sup>1</sup>, 陈仕吉<sup>2</sup>

(1.中国农业大学图书馆, 北京 100193; 2.杭州电子科技大学中国科教评价研究院, 杭州 310018)

**摘要:** [目的/意义]明确论文新颖性的概念及不同测度方法, 归纳国内外新颖性测度的不同类型及研究进展, 这对科学发展和科技规划具有重要启发。[方法/过程]对已有关于论文新颖性及其测度的中英文文献整理归纳, 从新颖性定义、新颖性测度不同方法如内容分析法和引文分析法及评价的角度出发形成综述。[结论/结果]现有方法多是从单一角度出发测度论文新颖性, 较少结合多种方法对新颖性进行测度, 未来发展方向应当结合定量和定性的方法全面分析。

**关键词:** 论文; 新颖性测度; 综述; 文献计量

中图分类号: G353.1

文献标识码: A

文章编号: 1002-1248 (2020) 10-0056-06

**引用本文:** 柴嘉琪, 陈仕吉. 论文新颖性测度研究综述[J]. 农业图书情报学报, 2020, 32(10): 56-61.

## Review on Paper Novelty Measurement

CHAI Jiaqi<sup>1</sup>, CHEN Shiji<sup>2</sup>

(1.Library, China Agricultural University, Beijing 100193; 2. CASEE, Hangzhou Dianzi University, Hangzhou 310018)

**Abstract:** [Purpose/Significance] This paper clarifies the concept of novelty, reviews different novelty measurements, and summarizes different types and research progress of novelty detection at home and abroad. It is of great significance to the development of science and scientific technological planning. [Method/Process] By summarizing existing relevant Chinese and English literature, this paper reviews the definitions of novelty, different novelty measurements such as content analysis and citation analysis, as well as their evaluation methods. [Result/Conclusion] Most of the existing methods measure the novelty from a single perspective, and few used combined methods. Therefore, both qualitative and quantitative methods should be taken into consideration to form a comprehensive analysis in the future.

**Keywords:** paper; novelty measurement; review; bibliometrics

收稿日期: 2020-03-11

基金项目: 国家社科基金项目“多视角下跨学科研究影响力综合评价研究”(20BTQ083)

作者简介: 柴嘉琪(1995-), 女, 硕士, 研究方向为科学计量学。陈仕吉(1974-), 男, 博士, 教授, 研究方向为科学计量学

## 1 引言

2016年5月, 中共中央、国务院印发《国家创新驱动发展战略纲要》, 进一步明确“把创新驱动发展作为国家的优先战略, 以科技创新为核心带动全面创新”的新战略要求。十九大报告中, 习近平总书记指出: “创新是引领发展的第一动力, 是建设现代化经济体系的战略支撑”。《国家中长期科学和技术发展规划纲要》提出, 要把提高自主创新能力应摆在全部科技工作的突出位置。近年来国家全面提出要坚持创新驱动, 推动产学研结合和技术成果转化, 强化对创新的激励和推动创新成果应用, 加大对新动力的扶持, 培育良好创新环境, 这带动了学术领域科研创新。

随着科学生产力提升, 科研活动发展态势良好, 论文作为科研活动的主要形式, 每天都在大量产出。在学术研究和传承的过程中, 往往是创新力强的论文更能起到承上启下的作用<sup>[1]</sup>。论文的创新力象征着学术影响力水平以及科研发展的方向, 而新颖性是论文创新力评价的一个重要方面。随着计量学的发展, 从定量的角度出发探索论文新颖性已经产生了很多尝试。但学术论文的内容复杂, 形式多样, 定量的方式很难全方位地覆盖所有角度, 这就要求我们不断思考探索更完善的论文新颖性的测度方法。此综述的意义在于探索不同类型方法如何探测论文新颖性, 有助于更好地利用不同方法, 从不同层面去测度论文新颖度, 对比分析不同方法测度新颖性的类型差异, 挖掘不同方法的优缺点, 为科技创新评价提供新的切入点, 便于科研工作者评估自身研究的新颖性以有效开展跨学科协作。

## 2 相关概念界定

创新是指突破旧的思维定势, 旧的常规戒律, 创新活动的关键点是“新”。创新行为可以分为3类: 无创新、渐进性创新、突破性创新。渐进性创新是逐步的、连续性创新行为, 突破性创新是对现有技术的颠覆, 是根本的、变革性创新行为<sup>[2]</sup>。新颖性和突破性既有区别, 又有联系, 新颖性是事物事前展现出的特征, 而突破是事后产生的影响。并非所有的突破都来自于新颖的研究, 但新颖性的研究有助于突破的产生, 突破性可能来自于一系列累积和增量的研究<sup>[3]</sup>。国外学者LEE对创新的理解是, 创新包含有用和新颖两方面<sup>[4]</sup>。HEINZE等学者将创新性研究概括为: 革命性新理论、发现新现象、提出和使用新方法、发明新仪器、从新角度整合现有理论<sup>[5]</sup>。

目前已有研究从不同角度对学术论文创新力的含义进行了探讨。周露阳认为论文创新性有3层基本含义: 与已有学术文献“不同”的论文; 不同在于“所涉学术领域知识或信息”的不同且这种不同的知识或信息须是“有价值”的<sup>[6]</sup>。学者FAGERBERG指出: “新颖性是对已有思想、能力、技巧、资源等的新组合”<sup>[7]</sup>。ZENG等认为, “组合过程可以看作是一个将与知识库中不同距离的知识联系起来的新颖性过程。”<sup>[8]</sup>。邢美凤等人认为新颖性测度是比较先前文件和现有文件间的相似度, 确认新文本内容中所含有的新成分<sup>[9]</sup>。因此, 本文借鉴ARTHUR对新颖性的定义, 从知识组合的角度出发, 新颖性可被认为是以前所未有的方式重新整合已有的知识<sup>[10]</sup>。创新性和新颖性的比较详见表1。

表1 创新性和新颖性的比较

项目	创新性	新颖性
概念	提出或完善已有理论或观点, 提出新方法或改进已有方法, 发现新数据或对已有成果进行总结概括等	从知识组合的角度出发, 新颖性是以前所未有的方式重新整合已有的知识
特点	有用性、新颖性、变革性、价值性	时间性、直接不同性
主体	一般为学者、高校、期刊等组织机构, 通过创新活动产生创新成果	一般为论文
评价	使用被引频次等指标衡量影响力并结合新颖性同时评价	仅使用相似度等指标测度文献之间的新颖程度

### 3 论文新颖性测度方法

对新颖性探索 (Novelty Detection) 的研究最早可以追溯到 1996 年 9 月, 是由美国国防部 DARPA (Defense Advanced Research Projects Agency) 发起的首次报道检测 (First Story Detection or New Event Detection), 在数据流中探寻首次讨论某个话题的新闻报道<sup>[1]</sup>。应用到信息检索领域则是文本信息检索领域最权威的国际性评测会议——文本检索会议 (Text Retrieval Conference, TREC), 在 2002 年召开的第 11 次会议新增了对文章内容新颖性追踪的项目<sup>[1]</sup>。自此之后, 国内外相关领域的研究者们开展了多种对文献新颖性的测评研究。

对于论文新颖性的评测, 主要使用的方法有定性评价和定量测度。从定性评价的角度出发, 同行评议作为一种重要方法, 在论文新颖性评测方面发挥了重要的作用。同行评议是指由期刊编辑或行业内资深的专家从定性的角度出发, 对论文的质量和水平以及内容的新颖程度等加以评判。同行评议中, 来自同一领域的学者, 结合自己的专业知识, 从各项研究成果中分析其创造性思路、代表性观点、标志性贡献等, 透过论文本身去辨识和评价论文中承载的知识内容的创新程度、创造性价值等<sup>[2]</sup>, 避免了由量化模式评价带来的结果同质化问题, 对论文新颖性进行了有效规范的评价。对外行人来说, 同行评议更能以质性的方式帮助研究人员认识研究成果, 但由于其结果与学者的个人认知直接挂钩, 因此其具有一定的主观性和非准确性。随着计算机技术的发展, 从数据科学的角度出发, 结合相关算法, 已为同行评议过程提供了一定程度上的支撑和参考。总的来说, 同行评议的重点在于评价主体的客观公正性和评价方法体系的科学合理性。

从定量测度的角度出发, 学界对论文新颖性的测度, 主要从两个层面出发, 一是基于内容分析的新颖性测度, 二是基于文献计量的新颖性测度。从不同角度出发探索论文新颖性各有优劣, 都在一定程度上量化了论文新颖度, 实现了对论文实质内容产生新颖性

成果的评测。以下, 本文将着重论述不同新颖性定量测度方法及其优劣势。

#### 3.1 基于内容分析的新颖性测度方法

在基于内容分析的已有研究中, 国外学者主要使用了向量空间模型、主题探测追踪等方法, 使用词语、语句、文本等不同指标对论文新颖性进行判断。

##### 3.1.1 使用向量空间模型的测度新颖性

目前来看, 新颖性测度方法中使用较为广泛的有基于向量空间模型的方法, 其中应用最为广泛的是向量夹角余弦值, 通过相似性判断论文新颖性, 文本之间相似性越大, 新颖性越小<sup>[3]</sup>。ZHANG<sup>[4]</sup>以向量空间模型为基础进行新颖性探测, 认为当前文本和以前本文之间的相似性值越大, 新颖性越小。逯万辉等基于 Doc2Vec 算法工具包计算文本语义相似度, 构建学术成果的主题新颖性测度算法<sup>[5]</sup>。ZHANG 等<sup>[6]</sup>提出基于词重叠度的新颖性判定方法, 通过查看其文本内容是否有所重叠, 重叠度越低表示文献越新颖。陈斯斯等认为词重叠法比逆文档频率量化法探测医学文献主题新颖性更优越合适<sup>[7]</sup>。还有如基于命名实体识别的文本新颖性检测, KUMARAN 等<sup>[8]</sup>通过结合文本分类和命名实体技术, 改进了基于向量空间的新颖性探测方法对文档的表达, 提高了新颖性的探测精度。以及基于 K 近邻的新颖性检测和使用聚类的方法等<sup>[9]</sup>都从不同侧面对论文新颖性进行测度, 较为客观地量化了论文的新颖性。 TSAI 等<sup>[10]</sup>应用文档—语句注释框架结构 (D2S, Document-to-Sentence) 方法进行新颖性探测, 先将文本分割成语句, 得到每句话的新颖性, 之后在固定阈值的基础上, 进一步研究计算单篇论文的具体新颖性分值。基于内容的文本分析技术能够较为客观与完整地展现文本内容, 具有一定的优势<sup>[20]</sup>。使用向量空间模型的方法对自然语言处理技术及深度学习等计算机技术要求较高。

##### 3.1.2 从主题探测角度出发测度新颖性

学界也有学者以寻找研究前沿的方式建立模型, 这在一定程度上量化了文献主题或学术成果的新颖性。学者使用主题词、关键词、词频等指标开展了如下研

究, RAJARAMAN 等<sup>[21]</sup>从神经网络角度出发, 提出了主题新颖性检测和跟踪算法。沈阳提出了基于关键词频度、使用的时间跨度及其他因素的文献创新度评价, 他认为创新频度越高, 其文本新颖性越大<sup>[22]</sup>。杨建林使用带时间戳关键词对逆文档频率 (Time-Stamped Inverse Document Frequency of Keyword Pair, KPTIDF) 量化学术论文的主题新颖度<sup>[1]</sup>。任海英等从论文的标题、摘要及关键词部分提取论文的主题词, 构建领域主题词共现网络<sup>[23]</sup>, 为每篇论文设计新颖组合率、中等组合率和常规组合率 3 个指标来表征论文新颖性。杨京等人从研究主题出发, 使用 Keygraph 算法抽取关键词, 后将论文主题与研究前沿进行相似度计算, 最后结合外在指标如期刊影响因子和 Altmetrics 等<sup>[24]</sup>对论文创新力进行综合评测。MENG 等采用词表的方法 (如基于 WordNet 的语义相似度计算方法) 对作者给出的关键词进行规范处理和语义扩展<sup>[25]</sup>, 计算文本相似度得出新颖性。使用词语的方法依赖于全面的语义词典, 其应用范围相对较为局限。此外, 还有学者使用了共词分析法, 通过统计一词组 (关键词或主题词) 两两在同一篇文章中出现的频次, 以计算这些词组之间的亲疏远近, 分析这些词组所代表的主题之间的结构变化情况<sup>[26]</sup>。许丹提出一种基于自然语言词对的量化指标—文档主题新颖度, 即含有某一共现自然语言词对的文献发表越早, 其新颖度越高<sup>[11]</sup>。吴菲菲等在探索研究前沿创新性时提出, 当研究中出现新的关键词且该关键词在近期被大量引用时, 预示着该关键词的创新特征明显, 他采用 Kleinberg 设计的跳跃检测算法 (Burst Detection Algorithm) 计算关键词的突现性, 将此作为研究前沿创新性的评价标准<sup>[27]</sup>, 这对论文新颖性的探索有一定的借鉴意义。从词角度出发探测论文新颖性, 在微观层面揭示了论文新颖性的表现, 也在一定程度上量化了论文新颖性, 但词语是否能真正代表论文的新颖性还有待探究。

### 3.2 基于引文分析的新颖性测度方法

新思想的基石往往在已有的知识中体现, 通过不同的方式整合和重构, 平衡新知识与传统知识对论文

新颖性及其产生的影响至关重要。影响更大的科学往往是建立在不同于传统知识组合的基础上, 同时又有新组合的加入。事实上, 已有研究中很多学者都将组合视为创新的核心。专利文献中引用的参考文献也被用于追踪科学创新的来源<sup>[28]</sup>。从网络结构角度来看, 新颖性也可以被理解为在科学网络中建立新的联系或桥接结构<sup>[29]</sup>。

从计量学角度出发, 不少研究通过分析文章的参考文献来测度论文新颖性<sup>[30]</sup>。每篇文章的参考文献中, 成对的组合可以作为评估新颖性产生的依据, 不同期刊可以看作不同领域的知识, 不同知识组合有可能产生新知识, 基于参考文献组合的新颖性测度方法更有可能体现学科知识真正意义上的交叉融合。2013 年, UZZI 等运用蒙特卡罗模型生成随机引文网络的方法对引文网络中引文组合的新颖性与常规性进行研究, 使用共同引用的期刊关系来确定任何一对引用的参考文献是否具有新颖性, 负 z-score 反映新颖的知识关系, 正 z-score 表示传统知识关系<sup>[31]</sup>, 这为学界测度新颖性给出了更为科学的方法。BOYACK 等改进了 UZZI 等度量组合新颖性的方法, 选用联合引用期刊对的 K50 统计数据代替标准分数 z-score, 在计算层面更加方便<sup>[32]</sup>。WANG 等也将期刊作为知识的组成部分, 通过检查参考文献期刊对是否从未在以前的出版物中出现过来判断论文新颖性<sup>[3]</sup>。LEE 等通过计算参考文献每一期刊对的相似程度 (Commonness) 来量化论文新颖性<sup>[4]</sup>。使用论文的知识来源 (如参考文献) 等信息作为测度论文新颖性也存在一些不同的研究结论, 如 TAHAMTAN 等研究表明, 一些已发表的研究可能对新发现很重要, 但未必在论文中引用; 引用的参考文献可能与论文新颖性有关, 但创意不一定受到他们的启发<sup>[33]</sup>。从引用结构方面来看, FOSTER 等用复杂网络描绘论文中知识的演化, 若一项研究能把两个子网络上的文献连通, 或能把两项未连通的研究连通, 这种组合就是新颖的<sup>[34]</sup>。此外, 跨学科性其中一个因素, 差异度, 对于新颖性产生也非常重要, 由于距离的存在, 参考文献差异度越高, 论文新颖性越强<sup>[35]</sup>。

## 4 结 语

除定性的同行评议外,目前对于论文新颖性的定量研究,一般有内容分析法和文献计量法,它们分别结合了不同的理论、工具,从不同维度测度了论文新颖性。

现有基于内容分析的新颖性测度方法在一定程度上探究了学术论文的新颖性。具体来说,它的优势在于使用词语、词频在揭示文章主要内容、表达主旨含义上更加准确、科学。劣势在于,第一,作者使用关键词时可能存在不规范的情况或语义问题,一个名词有多种含义,例如“苹果”,既可以指苹果公司,也可以是一种水果,这些不合理的现象会影响论文新颖性测度。第二,词语不能全面反映论文实质内容的新颖性,术语的变化很可能只是表象的创新,或是原词语由于社会环境的变迁以及技术进步或需求改变发展演变形成的替代词,例如从信息组织到知识组织的词语演变。由此可见,在测度论文新颖性前,应对关键词进行合理、规范化预处理,降低同近义词、上下位词对分析结果产生的影响。第三,基于关键词探测论文新颖性在时间上存在一定的滞后性,新的关键词达到规模,形成趋势需要时间的积累,以单篇论文为例,在文章发表后并不能立刻测度其新颖性。使用主题词表的方法亦存在此问题,虽然主题词表每年都在更新,但仍然滞后于科学研究的发展。新词汇从出现到被认可到认可需要时间的积累和专家学者的推荐,不能及时测度文献新颖性。

现有基于引文的新颖性测度方法的优势在于,它利用大量的引文数据,提升了测度结果的科学性,这是内容分析所做不到的;同时,在文章发表后可以立刻对其测度,更早识别出论文是否具有新颖性,发现论文的价值,便于对新颖性进行客观的测量。但基于引文的测度也有局限性,引用对论文实质内容的影响是间接的,从引文角度出发探测文章新颖性有可能偏离文章的实际内容,甚至无法从根本上判断论文的新颖性。

衡量论文的新颖性,最终讨论的还是其内容的新颖程度,未来研究若能将内容分析与引文分析两种方法结合使用,既注重实际内容的新颖程度,又利用了引用关系辅助比较,或许能够互相弥补不足,优化探测结果,较为全面地揭示学术论文的新颖性。除此之外,利用定量的方法分析论文的外在特征时,还需要结合定性的同行评议对其内容进行深层次、全面化的评价和揭示。

### 参考文献:

- [1] 杨建林,钱玲飞.基于关键词对逆文档频率的主题新颖度度量方法[J].情报理论与实践,2013,36(3):99-102.
- [2] 党倩娜,杨倩,刘永千.基于大数据方法的新兴技术新颖性测度[J].图书馆杂志,2019,38(4):91-100.
- [3] WANG J, VEUGELERS R, STEPHAN P E, et al. Bias against novelty in science: a cautionary tale for users of bibliometric indicators[J]. Research policy, 2017, 46(8): 1416-1436.
- [4] LEE Y, WALSH J P, WANG J, et al. Creativity in scientific teams: Unpacking novelty and impact [J]. Research policy, 2015, 44(3): 684-697.
- [5] 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.
- [6] 周露阳.论评审学术论文创新因素的指标体系[J].编辑学报,2006(1):68-70.
- [7] YEGROS-YEGROS A, RAFOLS I, D'ESTE P. Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity[J]. PloS one, 2015, 10(8).
- [8] ZENG A, SHEN Z, ZHOU J, et al. The science of science: From the perspective of complex systems[J]. Physics reports, 2017, 714: 1-73.
- [9] 邢美凤,过仕明.文本内容新颖性探测研究综述[J].情报科学,2011,29(7):1098-1103.
- [10] ARTHUR W B. The nature of technology: What it is and how it evolves[M]. New York: Simon and schuster, 2009.
- [11] 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):130-13.
- [12] 曾建勋.破除“SCI至上”需要弘扬“同行评议”[J].数字图书馆论坛,2020(2):1-2.

- [13] 许丹, 徐爽, 陈斯斯, 等. 主题词法和自然语言法探测文献主题新颖性对比分析[J]. 中华医学图书情报杂志, 2019, 28(1): 19-26.
- [14] ZHANG Y, TSAI F S. Chinese novelty mining [C]//Proceedings of the 2009 conference on empirical methods in natural language processing: Association for computational linguistics, 2009, 3: 1561-1570.
- [15] 逯万辉, 谭宗颖. 学术成果主题新颖性测度方法研究——基于 Doc2Vec 和 HMM 算法[J]. 数据分析与知识发现, 2018, 2(3): 22-29.
- [16] ZHANG M, SONG R, LIN C, et al. Expansion-based technologies in finding relevant and new information: The trec 2002: Novelty track experiments [J]. Nist special publication sp, 2003, 251: 586-590.
- [17] 陈斯斯, 董立平, 许丹, 等. 医学文献主题新颖性探测方法对比分析[J]. 中华医学图书情报杂志, 2018, 27(2): 20-25.
- [18] KUMARAN G, ALLAN J. Text classification and named entities for new event detection[C]//Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, 2004: 297-304.
- [19] TSAI F S, ZHANG Y. D2S: Document-to-sentence framework for novelty detection [J]. Knowledge and information systems, 2011, 29(2): 419-433.
- [20] 王颖. 科技文献内容语义描述模型研究[J]. 农业图书情报学报, 2020, 32(8): 12-24.
- [21] RAJARAMAN K, TAN A H. Topic detection, tracking, and trend analysis using self-organizing neural networks [C]//Pacific-Asia conference on knowledge discovery and data mining, Springer, Berlin, Heidelberg, 2001: 102-107.
- [22] 沈阳. 一种基于关键词的创新度评价方法[J]. 情报理论与实践, 2007(1): 125-127.
- [23] 任海英, 王德营, 王菲菲. 主题词组合新颖性与论文学术影响力的关系研究[J]. 图书情报工作, 2017, 61(9): 87-9.
- [24] 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.
- [25] MENG L, HUANG R, GU J. A review of semantic similarity measures in WordNet[J]. International journal of hybrid information technology, 2013, 6(1): 1-12.
- [26] 薛辰. 国际大数据研究论文的计量分析[J]. 现代情报, 2013, 33(9): 129-134, 139.
- [27] 吴菲菲, 杨梓, 黄鲁成. 基于创新性和学科交叉性的研究前沿探测模型——以智能材料领域研究前沿探测为例[J]. 科学学研究, 2015, 33(1): 11-20.
- [28] NAGAOKA S, YAMAUCHI I. The use of science for inventions and its identification: Patent level evidence matched with survey[C]//Research institute of economy, trade and industry (RIETI), 2015.
- [29] CHEN C, CHEN Y, HOROWITZ M, et al. Towards an explanatory and computational theory of scientific discovery[J]. Journal of informetrics, 2009, 3(3): 191-209.
- [30] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the association for information science and technology, 1973, 24(4): 265-269.
- [31] UZZI B, MUKHERJEE S, STRINGER M J, et al. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157): 468-472.
- [32] BOYACK K W, KLAVANS R. Atypical combinations are confounded by disciplinary effects[C]//Proceedings of the science and technology indicators conference 2014, Leiden, 64-70.
- [33] TAHAMTAN I, BORNMANN L. Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references?[J]. Journal of informetrics, 2018, 12(3): 906-930.
- [34] FOSTER J G, RZHETSKY A, EVANS J A, et al. Tradition and innovation in scientists' research strategies [J]. American sociological review, 2015, 80(5): 875-908.
- [35] FONTANA M, IORI M, MONTORBIO F, et al. A bridge over troubled water: Interdisciplinarity, novelty, and impact [R]. Università cattolica del sacro cuore, dipartimenti e istituti di scienze economiche (DISCE), 2018.