·科学论坛·

变革性研究的科学计量学特征与早期识别方法

杜 建[1*] 孙轶楠[1] 张 阳[2] 唐小利[1]

(1. 中国医学科学院医学信息研究所，北京 100020；2. 中国医学科学院北京协和医院，北京 100730)

[摘 要] 变革性研究是指挑战或颠覆原有研究范式，能够创造新范式或新领域的研究。通过数据分析以及案例分析揭示变革性研究在可能首先遭遇负面引用(或自引)，后期被专利引用初现变革潜力，随后被科学领域的一篇重要综述引用，从技术回归科学，开始受到科学界的认同，继而在科学界自己学科以外的其他多个学科领域产生影响。本文提出这类孕育重大创新突破的延迟承认类文献可能遵循“同学科科学共同体内争论式扩散——从科学向技术交叉扩散——从技术再次反馈到科学，并在不同学科间扩散”的发展路径。与热点跟风式成果相比，重大创新成果往往遭遇延迟承认，这是重大科技创新的基本规律。以生命科学领域为例，综合运用延迟承认指数和专利引文中的非专利文献(NPL)分析方法，初步提出一个识别中低被引论文(under-cited)中可能孕育潜在突破论文的思路。

[关键词] 变革性研究；引用延迟；非专利引文分析

变革性研究这一概念主要来源于美国国家科学基金会，是指挑战或颠覆原有研究范式，能够创造新范式或新研究领域的研究。这些研究挑战着现有的科学认知，或者能够通往科学的新前沿[1-2]。当前，“颠覆性”“变革性”成为科技创新领域的热点和前沿议题，全球范围内很多国家/地区的基金资助机构，包括美国国家科学基金会、美国国立卫生研究院、欧盟研究理事会、中国国家自然科学基金委员会，都在强调对变革性研究的资助[3]。变革性研究也受到了我国政府及科技政策制定者的高度关注[4-5]。“颠覆性技术创新”还写入了党的十九大报告。

在实施创新驱动发展战略的大背景下，我国经济向高质量发展将会对基础研究领域的创新提出更高的要求。其中一个重要的方面就是加快推进基础研究领域颠覆性创新的前瞻遴选和培育。如能早期识别、及时评出并及早部署具有颠覆性创新的基础研究，对于我国科技创新强国建设具有重要战略意义。本文首先讨论变革性研究、颠覆性技术、颠覆式创新的概念区分，然后通过案例研究尝试提出科学中潜在变革性研究的早期识别方法，开展实证研究。从科学学的角度，剖析重大科学发现的承认规律，以期为国家科技管理部门和基金资助机构早期识别、及早部署具有颠覆性创新的基础研究提供决策支持。

1 概念辨析

目前最常见的 3 个概念分别是变革性研究、颠覆性技术(Disruptive Technology)和颠覆式创新(Disruptive Innovation)，笔者认为它们分别是基础研究、技术创新和产品市场 3 个不同角度的概念，即分别对应着 Science、Technology 和 Innovation(表 1)。

我们认为，颠覆性技术和颠覆式创新从本质上是不可预测的(很难做到早期识别)，前者强调突然带来的技术优势，后者强调技术所带来的颠覆性效应。所谓“优势”和“效应”都是后显的。根据库恩的科学范式，科学研究可分为渐进性研究(Incremental Research)和变革性研究(Transformative Research)两类，或称常规性科学(Normal Science)和革命性科学(Revolutionary Science)[6]。前者是指在现有研

收稿日期：2018-10-30；修回日期：2018-10-31

* 通信作者，Email：du.jian@imicams.ac.cn

表 1　变革性研究、颠覆性技术、颠覆式创新概念区别

概念	角度	主要关注方	特点
变革性研究	基础研究	基金资助机构	具有颠覆性创新的基础研究，挑战或颠覆原有研究范式的研究。
颠覆性技术	技术研发	技术创新机构，如 DARPA	在国防领域应用更广。颠覆性技术的出现在于其突然带来的优势。可能会是技术组合的新方案，但是要能够解决现实问题，问题导向清晰，且潜在影响巨大。
颠覆式创新	产品市场	企业	强调技术所带来的颠覆性效应，而不是技术本身。颠覆式创新未必是突破性的（可以是现有技术），也未必是原创性的。产生颠覆式创新的技术，可以是基础性、原理性的新发现，也可能是现有技术跨领域、跨学科的创新性应用。

究范式下对已有研究的补充和发展，推动科学的累积式渐进；后者通常是对原有研究范式的颠覆，属于具有革命性的科学突破，促成科学革命的发生。渐进性研究与变革性研究并非二元独立关系，很多变革性研究建立在渐进性研究基础上，在后期表现出变革潜力[7]。脱离渐进性研究单独探讨变革性研究是不合理的。与颠覆性技术和颠覆式创新相比，挑战或颠覆原有研究范式的变革性研究可能在发表当时或初期就有“蛛丝马迹”，本研究试图寻找这类线索。

2　变革性研究在科学上的特征

美国国家科学委员会、国家科学基金会认为变革性研究具有以下特点：(1) 由挑战现状和颠覆传统研究范式的想法所驱动；(2) 会带来对传统科学理解的变革，甚至是颠覆；(3) 具有完全不同的研究路径；(4) 能够引领新的科学前沿，开拓新的领域。我国国家自然基金委政策专家认为基础研究领域的颠覆性创新具有以下特点[5]：(1) 思想驱动，具有偶然性；(2) 挑战传统，对现有认知进行颠覆，导致领域的革命性变化；(3) 初期难以达成共识，在同行评议中表现不佳；(4) 高风险性，成败概率不定，难以在前期计算投入产出效益；(5) 学科交叉，协同创新和综合交叉特征明显。

美国约翰霍普金斯大学和华盛顿大学的学者认为，革命性的发现往往来自基础科学，且严重依赖于非革命性的研究。革命性的发现可能是概念性的或技术性的，导致新领域的创造，并且除了它们出现的领域之外，还对许多领域产生持久的影响[8]。美国国家科学院《促进地理科学的变革性研究》报告通过回顾过去 65 年地理科学领域的变革性研究是如何出现并演化的，指出变革性研究可以从个人、团队以及各种知识来源中产生，包括一些比较老的或是长期被忽视的一些思想。目前的重要挑战是如何在变革性研究项目提出之时就能将其识别出来[9]。笔者认为，比较老的或是长期被忽视的一些思想反映的就是科学中的延迟承认，即睡美人文献。这类文献是指“发表后长期未被引或低被引，一段时期后突然高被引”的论文，是从科学计量学角度对科学社会学延迟承认现象的定量描述，其本质是超前性研究或变革性研究[10]。英国学者将诺贝尔奖成果作为变革性研究的代表，分析全球产生科学革命的国家和机构，表明美国是唯一一个大规模支持科学革命的国家。但欧洲、英国、中国的科技论文正在赶超美国（目前中国已超美国），科技论文总量反映的是常规性科学规模。如果能及时发现潜在变革性的研究，为少数的卓越科学家和科研机构提供差异化的基金资助和评估标准，则可能有助于科学革命的发生。呼吁需要为变革性研究或科学革命提供一套独特的科学计量学方法和指标[11]。

变革性研究在科学计量学方面具有哪些特征？本文将通过 *Faculty of 1000*、*Science* 和 *Nature* 较大规模数据分析和典型案例分析，尝试揭示潜在变革性研究的早期线索。

3　变革性研究在科学计量学上的特征

3.1　被引次数和同行评议在科研评价中的微观差异源于研究的渐进性 vs 变革性

笔者前期基于 Faculty of 1000 论文推荐分值和 Scopus 被引次数数据（28254 篇，发表时间 1999—2010 年，引文窗为 1999—2013 年），分析了不同研究类型对推荐分值和被引次数差异的影响[12]。根据推荐分值与被引次数百分位数分布，将论文分为 4 组：双高型、双低型、高推荐-低被引型和低推荐-高被引型。对照研究结果发现：(1) 标识为“新发现”“确认”“技术进步”“临床试验”“综述评论”和“系统综述/meta 分析”的论文得到了相对高的被引但却很少被同行推荐，多为“确认型研究”和“证据型研究”；(2) 标识为“有趣假设”“争议”“反驳/颠覆”“提供新药靶点”“能改变临床实践”的论文受到专家的

高度推荐但被引次数却相对较少，多为“变革性研究”和“转化型研究”。

可见，与渐进性研究相比，变革性研究往往低被引。该结论在胡小君和鲁索的研究中也得到验证。一些诺贝尔奖获得者的主要论文虽然原创性强，但并未获得与科学价值相匹配的被引次数。但是该论文的后续论文（第二、三代施引论文）却获得了极高的被引次数。这些“低引”的文献往往在科学发现的过程中扮演着“推动者”的角色，它们往往也是学科领域进步与发展的“基石”。这类奠基性研究提出了基本思想，但后续研究真正把该思想实现或转化，因此吸引了更多的引用[13]。我们前期对 2014 年诺贝尔化学奖的分析同样表明，首次打破物理学领域衍射极限经典范式的变革性研究的被引次数比后续在此基础上的拓展性研究的被引次数要低[14-15]。由此可见，起到奠基作用的变革性研究往往低被引，而由此引发的后续渐进性研究则往往高被引。

3.2 变革性研究多为睡美人文献且多具有技术属性与跨领域扩散特征

对典型睡美人文献案例的回顾性研究显示[16]，睡美人文献具有跨学科性、技术与应用属性以及变革突破性三大特征。睡美人文献反映的创新成果多是变革性研究，这类研究若遭遇延迟承认，可能存在跨领域唤醒机制，即在一个领域提出的创新思想可能在新的领域有了用武之地。变革性研究的典型特征——初期难以达成共识，在项目申请同行评议中表现不佳，也往往表现为关键性论文发表初期遭遇忽视或抵制。为识别睡美人文献或测度科学中的延迟承认，笔者提出一个新的用于识别睡美人文献的无参数指标——Bcp 指数，并应用该指标识别出 *Science* & *Nature* 上 1975—2005 年发表的若干睡美人文献，对排序前十文献内容分析表明，这些文献都是物理学、化学、生命科学或医学领域做出重大突破的文献，例如关于光分解水的经典文献、分子生物学的中心法则等，其中 3 篇是诺奖得主的文献[17]。

根据直观理解，人们往往认为延迟承认文献多倾向基础性、前沿性和理论性，提出的高深理论或概念超前于当时条件与认知水平，因此不被理解而遭遇延迟承认。因而延迟承认文献反映的似乎应该多为基础研究而非应用研究，似乎应多具有科学属性而非技术属性。但科学和技术是相互依存、相互促进的。科学研究是追求真理和知识创新的活动，并以其研究成果提升社会实践为目的。任何科学研究都需要真正的施惠人类社会才是其最终价值的体现[18]。笔者假设，既然延迟承认文献总是与重大科学发现相关联，而这样的科学发现之所以重要是因为它必定具有实践的属性。因此，延迟承认型文献在本质上可能更具有技术属性。

3.2.1 *Science* 和 *Nature* 延迟承认论文

对 *Science* & *Nature* 期刊在 1975—2005 年发表的 2 万篇被引 200 次以上（被引次数截至 2015 年底）的文献进行挖掘，识别延迟承认型文献。按照延迟承认指数（Bcp 指数）排序，将 Top 1% 视为延迟承认型文献，将 Bottom 1% 视为昙花一现型文献。通过对照研究分析这两组文献被专利引用情况。

表 2 延迟承认型文献和昙花一现型文献被专利引用情况

	Top1%	Bottom 1%	Total
论文	200	200	20 000
被专利引用的论文	98	70	13 729
被专利引用的论文所占比例	49.0	35.0	68.6
专利施引次数（按专利族统计）	3 361	494	109 859
篇均专利施引次数（按专利族统计）	34.3	7.1	8.0

注：论文被专利引用数据来源于 Lens.org，统计截至 2018 年 7 月 23 日。

表 2 可见，延迟承认型文献平均被 34 项专利技术引用，远高于昙花一现型文献（7 项）和全部文献的平均水平（8 项）。接下来，分析 200 篇延迟承认文献（发表年 1970—1994）的一代施引文献发现：（1）93% 的文献被最近 10 年（2008—2018，截至 2018 年 7 月 23 日）的 ESI 高被引论文所引用，即 20—45 年前的延迟承认论文对当前的研究前沿仍有影响；（2）每篇延迟承认文献的施引文献的 Web of Science 学科类别数平均值为 62.1，反映了这些遭遇延迟承认的早期研究至今对多学科领域产生了影响。

3.2.2 变革性研究案例：2014 年诺贝尔化学奖

在前期对 2014 年诺贝尔化学奖分析基础上，我们发现（图 1），Stefan W. Hell 被诺奖评审委员会认定为代表作（key publications）的两篇论文均为延迟承认型文献。这两篇文献（*OPT LETT*，1994，V19，P780 和 *APPL PHYS B*，1995，V60，P495）发表后立即被 Hell 本人的专利（US5731588，优先权年 1994，美国专利授权年 1998）引用。而这项美国专利仅引用了 2 篇非专利文献，恰好是这 2 篇代表作。两篇延迟承认文献被专利引用的时间早于唤

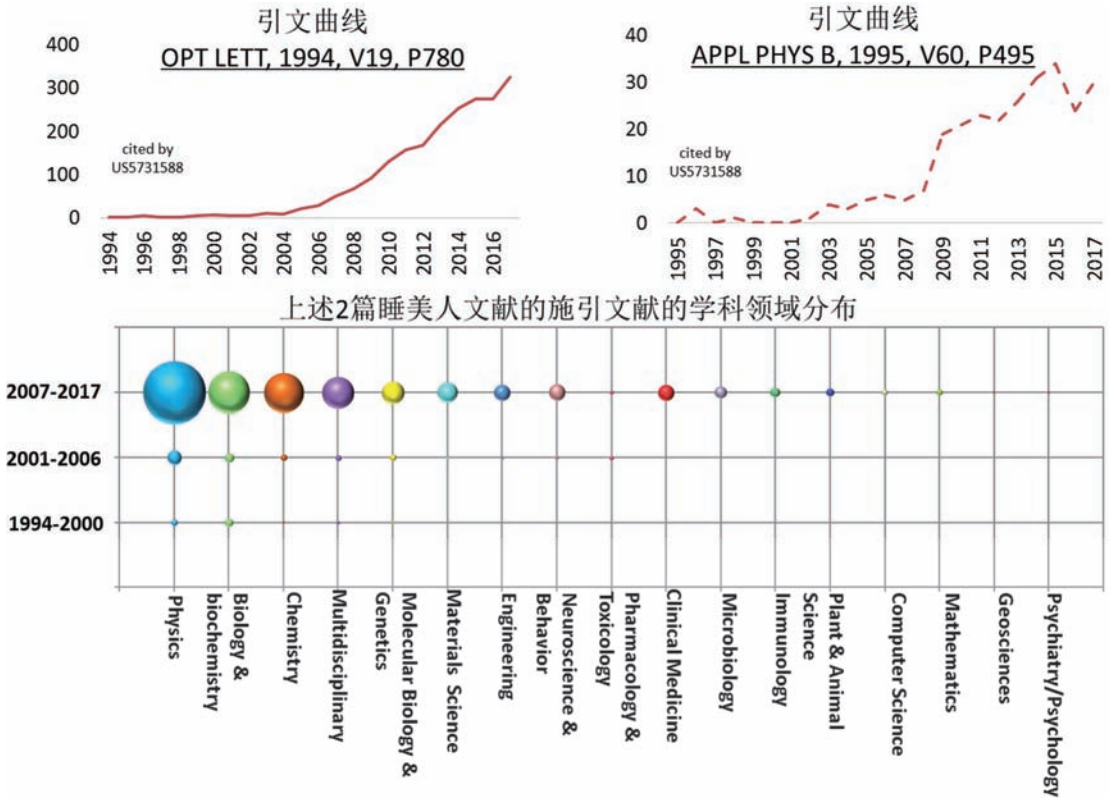

图 1　2014 年诺贝尔化学奖者得主 Stefan W. Hell 的变革性研究逐步受到承认的过程

醒年(2000 年)。Hell 本人 1994—1995 年接连发表的 2 篇文献均为延迟承认型文献，说明他的研究得到诺奖并非偶然成功，而是一个规划的很好(well-planned)的研究，这样的研究可能预示着未来的研究前沿。

将 2 篇延迟承认文献的一代施引文献按照发表时间划分为 3 个区间，可见该研究尽管打破物理学的经典范式，但创新扩散到了生物化学、分子生物学与遗传学、材料科学、工程学、临床医学、神经科学与行为学、免疫学等学科。高分辨率显微镜技术不断得到成功应用和发展，特别是在生物医学领域。

该论文标题和摘要术语表达本身就具有很强的变革性，如“propose a new type of…, we overcome the limit by …. In contrast to …, this method can…”。可见 Hell 本人坚信自己研究的创新价值。在其早期被引次数中，40%(12/30)施引为自引，Hell 对自己 1994 年的研究作出高度评价：“首次提出有望打破阿贝光学衍射极限的概念(the first viable concepts)”。可见，在标题或摘要中常出现上述术语的研究可能是潜在的变革性研究。通过文本识别潜在的变革性研究并追踪其转化应用状况，包括监测作者是否持续开展该主题的研究，是否从理论研究拓展到实践研究，实践中是否成功；论文发表之后是否有专利授权，论文是否早期被专利引用等，可能成为早期识别变革性研究的一个线索。

3.2.3　变革性研究案例：Smith 关于蛋白质空间的概念

另一个案例是一篇有证据支持的超前性研究，即 John Maynard Smith 的蛋白质空间的概念，论文发表于 1970 年的 *Nature* 杂志，题目为“自然选择与蛋白质空间”(Natural Selection and Concept of a Protein Space)，简称 Smith(1970)。该论文发表当年和第 2 年被引 3 次，但从引用情境上均为负面引用(表 3)。接下来的 12 年仅被引 6 次。1990s 被引次数开始增长，2005 年被引次数开始急剧增长，是一篇典型的睡美人文献。2004 年，Smith 先生去世时，*Nature* 杂志开辟纪念专栏，提到这篇文章提出的超前概念为现在大量应用的分子进化学的研究做了铺垫[①]。我们注意到 Smith 在论文中这样描述“Salisbury has argued that there is an apparent contradiction between two fundamental concepts of biology… Natural selection is therefore ineffective because it lacks the essential raw material — favorable mutations… I should like to look at the

① http://www.nature.com/nature/focus/maynardsmith/.

problem from a different point of view”。可见，该论文本身就有变革性研究的意味。结合表 3 可知，Smith 和 Salisbury 互相学术批判。

接下来，从施引论文和共被引论文两个角度寻找对 Smith(1970)被引次数增长起关键作用的文献，分别按照被引次数和共被引被引进行排序，在两个 Top10 列表中均出现了 2 篇论文。一是 1987 年发表于 *Journal of Theoretical Biology* 的“Towards a general-theory of adaptive walks on rugged landscapes”。这篇论文与 Smith(1970)的共被引次数最高，即：后续研究在引用 Smith(1970)时，总是同时引用此文。另一篇是 2005 年发表于 *Nature Review Genetics* 的一篇综述“Missense meanderings in sequence space: A biophysical view of protein evolution”。内容分析发现，Smith 首先提出了蛋白质空间(protein space)这一超前的概念，指的是蛋白质有许多种排列方式，然而有功能的却很少，犹如大海中的茫茫小岛。在分子进化过程中，有意义的突变只能从邻近的小岛到另一个小岛，即分子进化是一个逐步的过程。1970 年代，分子进化主要着眼于氨基酸序列(Sanger 测序于 1977 年出现)，因此相关的文章并不多，限制了 Smith 文章的引用。1980 年代，DNA 测序的应用使得分子进化可以以 DNA 序列为研究对象，大大促进了它的发展。1987 年这篇文章初步总结了适应性进化，引用了 Smith 的文章。进入 21 世纪，随着大规模测序技术的应用，能够用于分子进化的序列大大增加，分子进化领域迅速发展。2005 年综述系统的阐述了序列空间(sequence space)在分子进化中的作用，序列空间是在蛋白质空间基础上发展并包含蛋白质空间的，因此这两篇文章对 Smith 引用次数增加是有帮助的，第二篇文章的贡献更大，它让更多的人了解序列空间这个概念，及它的前身蛋白质空间这个理论。因此，这 2 篇论文均可视为唤醒 Smith(1970)的“王子”文献(Princes，简称 PR)。

表 3 Smith(1970)早期的 3 篇施引文献

标题	期刊	发表年	施引时的评论内容
Natural Selection versus Gene Uniqueness	Nature	1970	SMITH has compared protein evolution with a popular word game…I would point out, *however*, that it is *not sufficient* to… Smith gives the impression that he is *disagreeing with* Salisbury … In conclusion, after *examining* Smith's argument, Salisbury's contention still seems to stand…
Doubts about modern synthetic theory of evolution	American Biology Teacher	1971	My particular doubt has been published (Salisbury 1969), scientists have taken their *shots at* it (Smith, 1970), and it has been *defended* (Spetner, 1970).
Protein structure and properties	Journal of the American Oil Chemists' Society	1971	Obviously *only* an infinitesimal fraction of the latent variability is expressed in actual protein structures.

表 4 从施引论文和共被引论文两角度寻找对 Smith(1970)被引次数增长起关键作用的文献

	施引论文(Top citing papers)				共被引论文(Top Co-cited papers)			
	期刊	**被引次数**	共被引次数	年	期刊	被引次数	**共被引次数**	年
1	Protein Sci	1193	5	1995	**J. Theor Biol**	**607**	**55**	**1987**
2	Manage Sci	672	0	1997	Science	617	52	2006
3	**J Theor Biol**	**607**	**55**	**1987**	Proc Sixth Internat Congr Genetics, Ithaca New York	825	37	1932
4	Annu Rev Biochem	453	5	2010	Evolution	342	35	2005
5	**Nat Rev Genet**	**334**	**26**	**2005**	P Roy Soc B-Biol Sci	497	33	1994
6	Nat Rev Mol Cell Biol	326	22	2009	Nature	244	32	2007
7	Science	316	13	2006	P Natl Acad Sci USA	486	26	2006
8	J Theor Biol	305	6	1986	**Nat Rev Genet**	**334**	**26**	**2005**
9	Biol Cybern	282	6	1990	Evolution	245	26	1984
10	Microbiol Mol Biol Rev	256	1	2012	Nature	184	24	2006

被引次数统计时间：发表年至 2017 年 11 月 16 日。

表 5　引用 Smith(1970)的 2 篇专利文献(专利族)

标题	申请年	申请人	专利族数	授权(公开)年	优先权年	专利号	施引专利族数
Method For Producing Novel Dna Sequences With Biological Activity	Sep 30, 1994	Univ Washington (University)	2	Oct 20, 1998	Jul 17, 1986	US 5824469	64
Methods, Systems, And Software For Identifying Functional Bio-molecules	Mar 3, 2003	Maxygen Inc, et al. (industry)	24	Sep 12, 2003	Mar 1, 2002	WO 2003075129	50

被引次数统计时间：发表年至 2018 年 5 月 11 日，数据来源：Derwent Innovation Index。

值得注意的是，Smith(1970)这一非常基础性的研究被 2 项专利所引用，分别是 US5824469(优先权年为 1986)和 WO2003075129(优先权年为 2002)。前者为大学申请，后者为企业申请。两项专利的申请时间与上述 2 篇王子文献的发表时间很接近，且这两项专利目前的被引次数比较高，在 Derwent 专利数据库中分别被 64 和 50 个专利族引用。内容分析发现，Smith 提出的蛋白质空间概念对由分子进化而引发的技术发展也有一定的影响。其很重要的一个应用方向是 DNA、蛋白质等生物大分子的优化设计、筛选，例如专利 1 运用此概念插入随机核苷酸，表达并且筛选具有更佳生物活性的突变体；专利 2 则通过构建蛋白质变异体库，分析每一个变异体的差异位点，统计活性差异，从而设计出生物活性更好的蛋白质分子。另一方面，相关生物技术的发展也让更多的人了解、熟悉 Smith 提出的蛋白质概念。随着分子进化理论的发展，它对各个方面的影响也逐步加强。蛋白质空间这一理论也相应地被更多人接受和应用。

综上，Smith(1970)这一超前性/变革性研究在发表之初首先经历了负面批评式引用，1987 年同时被一项专利技术和一篇王子文献引用，随后被引次数呈现增长趋势。2003 年又被另一项专利技术引用，2005 年被一篇重要的综述引用，随后被引次数呈显著增长趋势。该研究属于进化生物学领域，2003—2017 年对分子生物学与遗传学、化学、植物学与动物学、物理学、微生物学、要学、免疫学、临床医学等领域均产生了影响(图 2)。

3.3　重大创新突破研究的科学计量学特征

上述研究表明，重大创新突破可能具有四大科学计量学特征：一是与热点跟风式研究(多为渐进性研究)相比往往遭遇延迟承认，表现为引用延迟；二是与后续拓展性研究(多为渐进性研究)相比不一定高被引，往往呈现中低被引特征；三是往往具有科学-技术交叉特征，被专利引用特征明显。四是具有跨学科领域创新扩散特征，往往被自身学科之外的领域所引用。变革性研究早期可能先遭遇负面引用(或多为自我引用)，后期被专利引用初现变革潜力，随后被科学领域的一篇重要综述引用，从技术回归科学，开始受到科学界的认同，随后在科学界自己学科以外的其他多个学科领域产生影响。我们发现，睡美人文献如果被专利引用，可能会对其唤醒起到关键作用，同时可意味着超前性研究被理解或变革性研究开始被认识和承认。我们认为，非高被引、引用延迟和被专利引用可视为变革性研究的早期识别信号。

接下来联合采用引用延迟分析和专利引文中的非专利文献分析尝试识别生命科学领域遭遇延迟承认的中低被引文献中潜在的变革性研究。

4　实证研究：挖掘被专利引用过的中低被引延迟承认类论文

4.1　方法与资料

延迟承认文献不是一个 Yes/No 的概念，而是一个与时间相关的程度概念(time-dependent continuous phenomenon)，正如“高被引论文”的定义。为此，借鉴 ESI 高被引论文“同年同学科领域被引次数排名前 1%”的定义，延迟承认文献的识别也应考虑发表年和学科领域的问题。为此，实证研究将以 2003 年生命科学领域自然指数期刊上发表的论文为样本。选择 2003 年是考虑到至 2017 年有 15 年的引文窗，限定为自然指数(Nature Index)期刊是考虑到从较为优质的论文中识别。

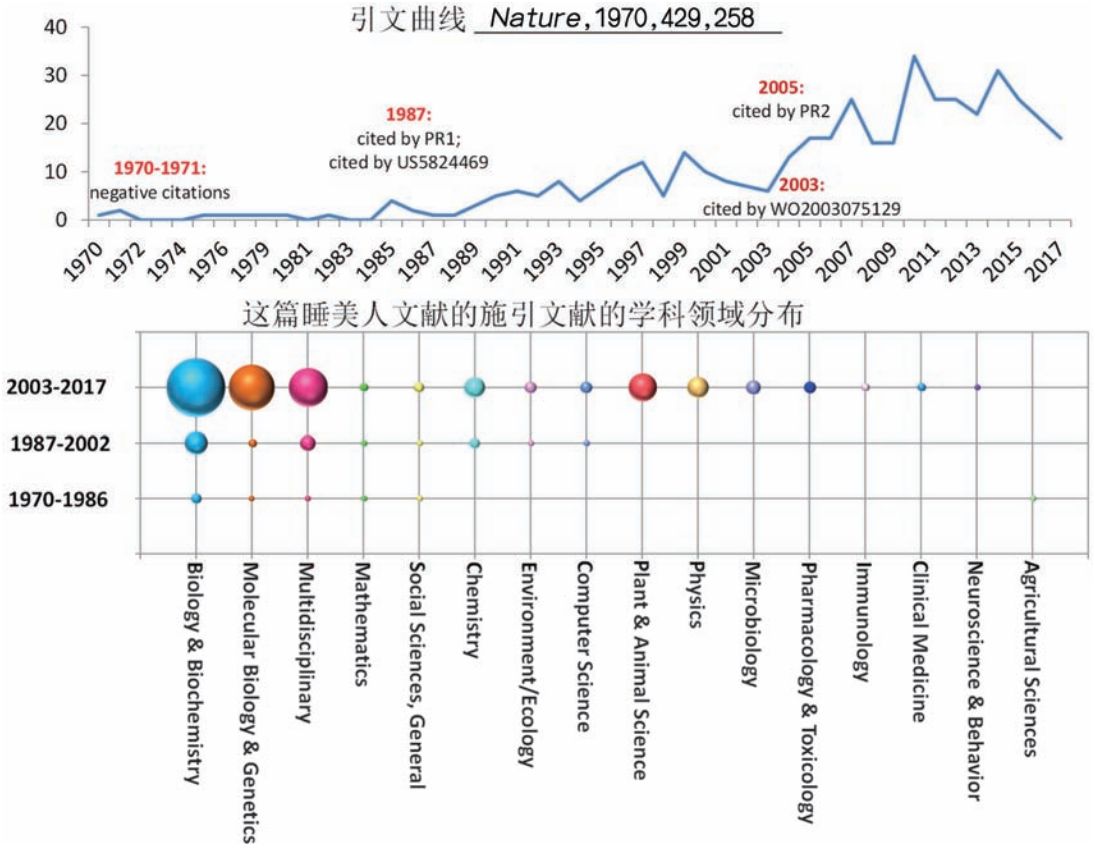

图 2 Smith(1970)关于蛋白质空间的超前性研究逐步受到承认的过程

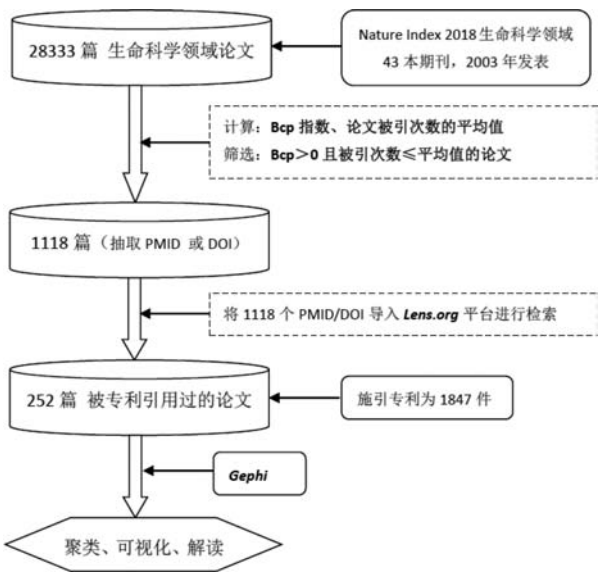

图 3 生命科学领域遭遇延迟承认的中低被引文献中潜在变革性研究的识别过程

自然指数 2018 生命科学领域的 43 本期刊 2003 年发表的全部文献共计 28 333 篇(检索时间:2018 年 7 月 29 日)。以期刊为单元,计算每本期刊上每篇论文的 Bcp 指数和整本期刊论文被引次数的平均值。依据 Bcp 指数和被引次数平均值,筛选出每本期刊上 Bcp>0(表示总体上延迟承认)且被引次数≤平均值的论文。最终筛查出符合条件的 1 118 篇论文,并抽取其 PMID 号(若无 PMID 则用 DOI 代替)。将 1 118 个 PMID/DOI 导入 *Lens.org* 平台进行检索,结果 252 篇论文被专利引用,施引专利为 1 847 件(图 3)。

4.2 分析结果

经过数据统计与分析,252 篇论文被专利共同引用次数≥1 次的论文共有 11 对。将论文-专利共被引矩阵通过聚类得到 9 个论文簇,邀请领域专家判读论文簇中的论文内容,根据论文判读结果,命名恰当的概念名称(图 4)。

将这 9 个主题与 2003—2017 年以来的 *Science* 十大科学突破和诺贝尔奖相匹配,发现每个主题至少与两个 *Science* 年度突破直接相关(表 4)。然后,我们对每个主题的创新性和突破性进行详细述评。其中免疫治疗获得了 2018 年诺贝尔医学奖。

主题 1。脊髓损伤后机体功能的改善一直是热点研究领域,早期研究更多关注于神经肌肉之间的结构以及二者如何联结,随着神经解剖的逐步明晰,并且伴随着生物材料、干细胞、神经科学、人工智能等的兴起,以及“再生医学与组织工程学”概念的扩

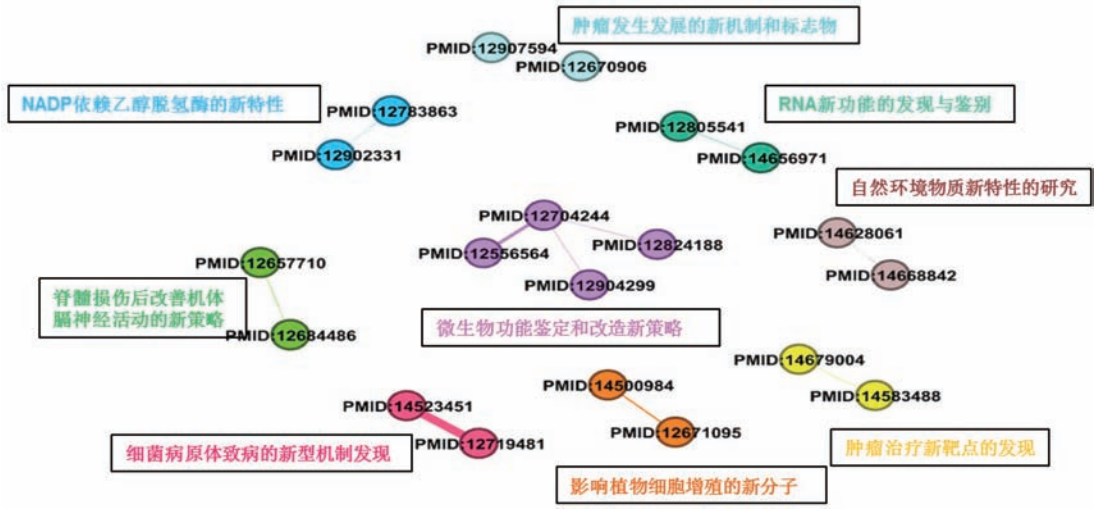

图 4　论文-专利共被引聚类图

表 6　9 个聚类主题与 *Science* 年度十大科学突破匹配(2003～2017 年)

	识别的主题名称	与 Science 年度十大科学突破的匹配(2003～2017)	相关诺贝尔奖
1	脊髓损伤后改善机体膈神经活动的新策略	操纵记忆(2014)；仿人脑电脑芯片(2014)；睡眠机制(2013)；大脑/机器界面(2012)；大脑记忆关键中心(2007)；制造记忆(2006)；精神分裂症根源(2005)	2004(嗅觉受体和嗅觉系统)；2014(大脑定位系统)
2	微生物功能鉴定和改造新策略	肠道微生物菌落研究(2011)；我们的微生物，我们的健康(2013)	
3	NADP 依赖乙醇脱氢酶的新特性	计算机设计蛋白(2016)；X 射线激光解开蛋白质的结构(2012)；光合蛋白结构(2011)；观察工作中的蛋白质(2008)；发现人体蛋白受体结构(2007)；发现看门蛋白质“近照”(2005)	2012(G 蛋白偶联受体的研究)
4	肿瘤发生发展的新机制和标志物	外显子测序找到癌症基因(2010)；新的致癌基因(2008)	
5	影响植物细胞增殖的新分子	植物 ABA 受体(2009)；植物开花之谜(2005)	
6	细菌病原体致病的新型机制发现	研制埃博拉疫苗(2015)；结构生物学指导疫苗设计(2013)；疟疾疫苗(2011)；HIV 预防(2010)；艾滋病防治(2011)	2005(幽门螺杆菌)；2008(HPV、HIV)；2015(疟疾、丝虫病)
7	RNA 新功能的发现与鉴别	RNA 重编程(2010)；PiwiRNA(2006)；基因组非编码区(2004)；RNA 功能的多样性(2003)	2016 年(RNA 干扰)
8	自然环境物质新特性的研究	270 万年前地球大气(2017)；钙钛矿型太阳能电池(2013)；日本“隼鸟号”小行星取样(2011)；发现宇宙最原始气体云(2011)；新的人工合成沸石(2011)；石墨烯材料研究(2009)；再生能源——可燃水(2008)；高温超导材料(2008)；过渡金属氧化物研究(2007)	2010(石墨烯)
9	肿瘤治疗新靶点的发现	广谱抗癌药(2017)；T 细胞同时具有两种功能(2007)；癌症免疫疗法(2013)；抗血管生成治疗肿瘤(2003)	2018(免疫治疗)

展，脊髓损伤的再生与修复进入到了一个新阶段，不单纯只注重于结构恢复，更关注于功能恢复以及病人远期预后。

主题 2。早期微生物研究集中于对其单一基因和代谢产物功能的鉴定；在应用上主要关注“发酵工程”，即采用现代工程技术手段，利用微生物的某些特定功能，为人类生产有用的产品，或直接把微生物应用于工业生产过程的一种新技术。而近十年来，随着下一代测序技术、代谢组学和基因编辑技术等的快速发展，以及转化医学观念的日益深入，微生物

领域的研究走向了多组学(如宏基因组、代谢组、蛋白质组)结合来系统了解人体与微生物功能之间的联系。而通过基因编辑技术(CRISPR/Cas9),可以快速定向改造微生物实现大规模应用。

主题 3。对蛋白质功能的鉴定是后基因组时代的一大迫切任务,随着冷冻电镜、代谢组学等的兴起,代谢通路中催化酶功能的研究更加深入,酶的结构以及代谢产物之间的联系逐步被揭示,推动了"酶工程"的快速进步。

主题 4。肿瘤发生、分化以及转移的研究始终是一大方向。近年来,由于各种组学技术、成像技术、以及单细胞测序技术等的发展,研究层次从以往的单一基因向网络集成发展。二代测序以及分子靶向药的应用加快了"个体化医学"的进程。

主题 5。对动植物个体的发育以及如何细胞增殖、修复的理解对于"发育生物学"研究至关重要。以往关注于单个基因单一时段在信号通路中的作用,但随着长程动态活体成像技术、示踪技术、测序技术等的发展,对生命体发育以及细胞增殖全谱的研究成为了可能。

主题 6。随着人类社会的进步以及生存空间的拓展,各种病原微生物传播的机率大大增加,发现其致病机制,并采用药物、疫苗等进行防治是重要研究课题。近年来,随着"微生物与微生物组学"概念的兴起,新型病原菌的快速诊断、预警预防有了质的飞越。

主题 7。RNA 功能的研究近二十年来的一大热点,对深入认识"中心法则"非常重要。尤其是既往认为无用的非编码的 RNA 的再发现,刷新了以往的认识。作为 DNA 与蛋白质之间的关键桥梁,未来 RNA 生物学仍会成为一大重要研究方向。

主题 8。如何利用自然资源、适应环境与改造环境是人类始终追逐的方向。研究自然资源的新特性以及发展新型的人工智能化材料,可以让人类在未来更好地适应自然环境。

主题 9。寻找肿瘤的新靶点来干预和治疗肿瘤是该领域的重要方向。近年来,随着精准医学概念的兴起,个体化靶向治疗日益深入人心。此外,免疫治疗在肿瘤领域不断取得重大突破,采用"联合免疫疗法"治疗肿瘤已成为一大趋势,将肿瘤变成慢性病进行管理成为了可能。

5 讨论与研究展望

本文通过延迟承认指数高和被专利引用两个指标识别出处于沉睡-唤醒、科学-技术的交叉处的研究内容,通过领域专家判读,这些研究主题均与 *Science* 年度十大科学突破甚至诺贝尔奖直接相关,验证了我们所提出的变革性研究的早期科学计量学线索的适用性。从目前积累的生物医学领域案例来看,睡美人文献多具有潜在的技术与应用属性,多与超前性技术或变革性技术有关。例如超分辨率荧光显微镜、高压氢气可治疗癌症、经皮腔内血管成型术等均是重大技术(手段)发明。睡美人文献概念提出者,荷兰莱顿大学 van Raan 最近研究表明[19],越是最新发表的睡美人文献,被专利首次引用的时间和睡美人文献发表时间之间的时滞越短。他认为,越是最新的睡美人文献,越倾向于被专利唤醒,强调了睡美人文献与变革性技术的相关性。重大创新突破可能先在技术领域被发现或被认可,然后再反馈到科学领域被认可,这也就解释了为什么睡美人文献在"睡眠期"(极少被科学论文引用)首先会被专利引用。因此,我们需要重视那些尽管在科学界被引很少,但却已被专利引用的论文,这类研究极有可能具有潜在变革性。

除延迟承认文献思路外,目前仅有少量研究在探索变革性研究的识别方法。一种观点是在传统研究范式下代表性成果形成的引文路径(引文链)中,变革性研究的出现表现为引发传统研究范式引文链的"破裂"。故台湾大学学者提出用"破裂分数"(disruption score)识别物理学、计算机科学和生物医学领域的变革性研究[20-21]。另一种观点是引用内容分析或施引语句分析方法被用来识别科学发现[22],或评价研究价值,特别是研究产生的学术传承效应[23]。两者的共同思路都是利用施引文献中的评论性引用语句来识别参考文献的内在价值。笔者认为,从延迟承认文献角度识别变革性研究是基于时间维度的,核心思想是寻找沉睡-唤醒的 interface 上的规律;而从引文网络既有路径发生突变或破裂的角度是基于空间维度的,体现了研究方向的变迁,核心是寻找研究方向-研究方向的 interface 上的规律,这里的研究方向可能是科学概念,也可能是技术点。如何更系统性的早期识别变革性研究,仍需要从时间和空间两个维度继续深入

探索研究。

下一步，拟继续采用引用文本分析识别变革性研究。首先构建显性的变革性研究术语集，即标题/摘要中出现作者描述的变革性研究的术语（如“disagree”，“contradict”，“inconsistent”，“dispute”等）的论文，以生物医学某主题领域的论文为案例，采用 PubMed Central 全文文本尝试探索基于引用内容分析用来识别变革性研究的方法，即利用施引文献中的评论性引用语句来识别被引文献研究内容的变革性。以此形成作者-同行对该研究变革性的双重认同，即不仅作者认为该研究具有变革性，同时也受到了同行基于评论性引用的认可。

致谢　本文获得中国科学青年人才托举工程（编号：2017QNRC001），国家自然科学基金（编号：71603280），中国医学科学院医学与健康科技创新工程（编号：2016-I2M-3-018）项目资助。

参 考 文 献

[1] 梁正，邓兴华，洪一晨. 从变革性研究到变革性创新：概念演变与政策启示. 科学与社会，2017，7(3)：94—106.

[2] Trevors JT，Pollack GH，Jr SM，et al. Transformative research：definitions，approaches and consequences. Theory in Biosciences，2012，131(2)：117—123.

[3] Sen A. Island＋Bridge：how transformative innovation is organized in the federal government. Science & Public Policy，2017.

[4] 郑永和，陈淮. 美国国家科学基金会加强支持变革性研究考察. 中国基础科学，2008，10(4)：39—42.

[5] 杨卫，郑永和，董超. 如何评审具有颠覆性创新的基础研究. 中国科学基金，2017(4)：313—315.

[6] Kuhn TS. 1970. The structure of scientific revolutions. University of Chicago Press，Chicago，IL.

[7] Gravem SA，Bachhuber SM，Fulton-Bennett HK，et al. Transformative Research Is Not Easily Predicted. Trends in Ecology & Evolution，2017.

[8] Casadevall A，Fang FC. Revolutionary Science. Mbio，2016，7(2)：e00158.

[9] National Academies of Sciences，Engineering，and Medicine. 2016. Fostering Transformative Research in the Geographical Sciences. Washington，DC：The National Academies Press. https://doi.org/10.17226/21881.

[10] 杜建. “睡美人”文献的识别方法与唤醒机制研究. 南京大学，2017.

[11] Charlton BG. Scientometric identification of elite ‘revolutionary science’ research institutions by analysis of trends in Nobel prizes 1947—2006. Medical Hypotheses，2007，68(5)：931.

[12] Du J，Tang X，Wu Y. The effects of research level and article type on the differences between citation metrics and F1000 recommendations. Journal of the Association for Information Science and Technology，2016，67(12)：3008—3021.

[13] Hu X，Rousseau R. Scientific influence is not always visible：The phenomenon of under-cited influential publications. Journal of Informetrics，2016，10(4)：1079—1091.

[14] 杜建，武夷山. 基于被引速率指标识别睡美人文献及其“王子”——以 2014 年诺贝尔化学奖得主 Stefan Hell 的睡美人文献为例. 情报学报，2015(5)：508—521.

[15] Du J，Wu YS. A Bibliometric Framework for Identifying “Princes” Who Wake up the “Sleeping Beauty” in Challenge-type Scientific Discoveries. Journal of Data and Information Science，2016，9(1)：50—68.

[16] 杜建，武夷山. 睡美人文献的重要特征、预测线索与科技政策启示. 科学学研究，2018，36(9)：1558—1565.

[17] Du J，Wu Y. A parameter-free index for identifying under-cited sleeping beauties in science. Scientometrics，2018：1—13.

[18] 董尔丹，胡海，张俊. 学术评价应更科学. 科学通报，2014，59(1)：96—106.

[19] Van AR，Winnink JJ. Do younger Sleeping Beauties prefer a technological prince? Scientometrics，2018，114(2)：701—717.

[20] Huang YH，Hsu CN，Lerman K. Identifying Transformative Scientific Research [C]//Data Mining (ICDM)，2013 IEEE 13th International Conference on. IEEE，2013：291—300.

[21] Huang YH，Ko MT，Hsu CN. Identifying Transformative Research in Biomedical Sciences [M]//Technologies and Applications of Artificial Intelligence. Springer International Publishing，2014：188—197.

[22] Small H，Tseng H，Patek M. Discovering discoveries：Identifying biomedical discoveries using citation contexts. Journal of Informetrics，2017，11(1)：46—62.

[23] 尚海茹，冯长根，孙良. 用学术影响力评价学术论文——兼论关于学术传承效应和长期引用的两个新指标. 科学通报，2016，61(26)：2853.

Characterizing and detecting the early scientometric signs of the potential transformative research

Du Jian[1] Sun Yinan[1] Zhang Yang[2] Tang Xiaoli[1]

(1. *Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing* 100020;
2. *Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing* 100730)

Abstract In this paper, we try to find the forecasting markers which will be of benefit to identifying potential sleeping beauty publications earlier and accurately. Using articles published between 1970 and 2005 in *Science* and *Nature*, combined with two typical cases of transformative research and ahead of time discoveries, we found that the non-patent literature (NPL) cited by patents and the interdisciplinary citations may provide insight to the awakening of sleeping publications, which means that the ahead of time discoveries get understood, or the transformative potential of research is recognized. It appears that the sleeping beauties firstly encounter negative citations and/or self-citations, and then patent citations and then interdisciplinary citations and finally get widely recognized. We propose to combine citation delay analysis with patent & NPL analysis to identify potential ahead-of-time and transformative research, especially the older and long-ignored ideas but now at both the sleeping-awakening interface and science-technology interface. These ideas and research topics may be the potential origin of transformative research, according to a recent report (National Academies of Sciences and Medicine 2016).

Key words Transformative Research; Citation Delay; Patent & NPL Analysis