● 张 琪[1,2]，马瑞敏[1]，杨 杰[2,3]，王左戎[2,3]，邓三鸿[2,3]
（1.山西大学经济与管理学院，山西 太原 030006；2.数据工程与知识服务重点实验室（南京大学），江苏 南京 210023；3.南京大学信息管理学院，江苏 南京 210023）

融合颠覆性与巩固性引文特征识别领域突破性论文*
——基于可解释性机器学习方法

摘 要：［目的/意义］识别突破性论文对于科技创新的前瞻性布局和科研管理具有重要意义。［方法/过程］基于可解释性机器学习方法，在作者影响力、被引频次等传统特征的基础上引入颠覆性与巩固性引文两种衡量创新性的特征指标，通过一系列实验对比XGBoost、TabNet和GcForest三种基础模型、不同超参数组合和不同时间窗口，从而获得最优领域突破性论文识别模型，并进一步采用SHAP方法评估各特征指标尤其是不同创新性特征指标在最优模型中的贡献程度。［结果/结论］①XGBoost在医学、物理和化学领域均表现最佳，最高*F*1值分别达到0.9482、0.9331和0.8061，能够较为有效地识别不同领域的突破性论文；②对于不同领域而言，模型均在引文窗口为10年时达到最佳识别效果，窗口缩小时识别效果显著下降，但下降幅度因学科而异；③不同领域的突破性论文识别中各特征指标的重要程度存在差异，但整体而言颠覆性创新特征均显示出较高的重要性，不同领域的巩固性创新特征的重要性程度则存在差异。上述发现可为科学评价与政策制定提供方法支撑与实证依据。［创新/局限］本研究创新点主要在于引入颠覆性和巩固性引文作为两种创新性类型的量化指标从而构建领域突破性论文识别模型，但未来仍需在更多学科情境下深入探讨不同创新性特征的重要程度及作用机制。

关键词：突破性论文；可解释性机器学习；自动识别；颠覆性指数

DOI：10.16353/j.cnki.1000-7490.2026.02.014

引用格式：张琪，马瑞敏，杨杰，王左戎，邓三鸿．融合颠覆性与巩固性引文特征识别领域突破性论文——基于可解释性机器学习方法［J］．情报理论与实践，2026，49（2）：126-133，81.

Fusion Disruptive and Consolidative Citations Feature Identify Breakthrough Papers: An Interpretable Machine Learning Approach

Zhang Qi[1,2], Ma Ruimin[1], Yang Jie[2,3], Wang Zuorong[2,3], Deng Sanhong[2,3]
(1. School of Economics and Management, Shanxi University, Shanxi Taiyuan 030006;
2. Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Jiangsu Nanjing 210023;
3. School of Information Management, Nanjing University, Jiangsu Nanjing 210023)

Abstract: [Purpose/significance] Early identification of breakthrough papers holds great significance for strategic planning in scientific and technological innovation and for effective research management. [Method/process] This study employs interpretable machine learning techniques, introducing two innovation-related features—disruptive citations and consolidating citations—alongside traditional features such as Author Impact and Citation Count. We conducted a series of experiments comparing three base models (XGBoost, TabNet, and GcForest) across various hyperparameter settings and citation time windows to construct the optimal domain-specific breakthrough-paper detection model. Subsequently, we applied SHAP (SHapley Additive exPlanations) to quantify the contribution of each feature, with a particular focus on the innovation-related metrics. [Result/conclusion] Our findings reveal that: ①XGBoost achieved the best performance in identifying breakthrough papers in medicine, physics, and chemistry, attaining peak *F*1-scores of 0.9482, 0.9331, and 0.8061, respectively, demonstrating robust cross-domain effectiveness. ②Across all domains, the optimal detection performance occurred with a 10-year citation window; shortening this window led to significant performance declines, although the extent varied by discipline. ③Feature importance varied across fields, but disruptive citation metrics consistently ranked highly in all domains, while the significance of consolidating citation metrics varied notably by discipline. These insights offer methodological support and empirical evidence for scientific evaluation and policymaking. [Innovation/limitation] This study's innovation lies in the integration of disruptive and consolidating citation metrics as quantitative measures of innovation for detecting breakthrough papers. However, further research across a wider range of academic disciplines is required to explore

* 本文为国家社会科学基金项目“基于证据的学者高水平论文代表作量化遴选研究”的成果之一，项目编号：22BTQ060。

the commonalities，differences，and mechanisms underlying feature importance in diverse contexts.

Keywords：breakthrough papers；interpretable machine learning；automated identification；disruptive index

0 引言

突破性论文（Breakthrough Paper）通常是指有较高创新性，对研究领域有相当大影响的科学著作[1-3]。人们相当广泛地一致认为诺贝尔奖（以下简称“诺奖”）代表了所授予学科的最高荣誉，因此在相关研究中通常将诺奖论文视为领域突破性论文[4-5]。在2035年建成科技强国的战略目标下，尽早地识别突破性论文对于科技创新的前瞻性布局和科研管理而言具有重要的意义[1]。

然而，传统依赖专家评议的推选方式对专家自身的研究范围提出了极高的要求，且时间周期长。而从指标计算的角度来看，已有研究表明，作者影响力、被引量、颠覆性指数等单一指标难以识别突破性论文[4]。创新是对科研论文学术价值及发表价值进行评价的重要因素，将学术论文的创新性指标纳入突破性论文识别中具有重要价值[6]。T. S. Kuhn[7]认为科研中存在两类创新：一是建立在前序研究的基础上并对其进行了增强；二是挑战现有范式并有可能颠覆现有领域。基于这一理论，R. J. Funk等[8]提出了衡量专利颠覆性的指标，并在后续被推广至科学文献计量领域。在此基础上，Yang等[9]将单篇论文的影响力划分为颠覆性引用和巩固性引用，形成双重创新性指标框架。然而，两种创新性特征指标是否能够提升不同领域的突破性论文识别效果及其重要程度仍然未知。

因此，本文在作者、参考文献等传统指标的基础上引入了颠覆性和巩固性引文特征分别衡量科研论文的颠覆性创新和巩固性创新，从而构建包含11个特征变量的指标体系。指标的重要性评估必须建立在良好模型效果的基础上，为此本文对比机器学习模型XGBoost（eXtreme Gradient Boosting）和两种深度学习模型Tabular及GcForest（Multi-Grained Cascade Forest）在突破性论文识别上的效果，层层递进地探索并回答了以下三个问题：①不同机器学习和深度学习模型在突破性论文识别上的效果如何？②不同时间窗口对突破性论文识别的效果是否有影响？③在获得最优模型的基础上最终探索不同特征变量，尤其是颠覆性引文量和巩固性引文量，对突破性论文识别的重要程度如何？

1 相关研究

本文基于可解释性机器学习的方法，融合学术论文的创新性特征构建多领域突破性论文识别模型，并进一步探讨不同特征指标尤其是创新性特征的重要程度。为此，本文从三方面回顾相关研究：首先梳理突破性论文的操作性界定及影响因素相关研究，然后比较不同突破性论文识别方法，最后聚焦于学术论文的创新性类型及其量化指标。

1.1 突破性论文及其影响因素

突破性论文是指对研究领域有相当大影响的科学著作，但其实际界定标准存在一定争议。以往研究多采用被引次数的阈值来界定（如前1%、3%或10%）[10]。但是，高被引论文并不一定是突破性论文[11-12]，而具有奠基性的突破性论文的研究往往比后续的拓展性研究被引次数低[13]。相较而言，诺奖因其在科学共同体中的极高影响力，相对应的“诺奖论文”常被视为突破性论文的典型代表，但也面临数据难以搜集、样本不均衡、难以识别等困难。为此，学界尝试以诺奖论文为中心构建突破性论文数据集。如Li等[14]结合诺奖官方网站与诺奖得主传记材料收集物理、化学和医学获奖论文，强化了数据集的可信性。在此基础上，相关学者采用统计学方法分析了作者特征（如职业成就、年龄、国籍）[15]和论文特征（如期刊影响力、被引次数）[16]对突破性成果的影响作用。这些研究不仅强化了突破性论文数据集的代表性与可信性，也为本文的指标选择提供了理论支撑。

1.2 突破性论文识别方法

在识别方法方面，现有方法大致分为计量指标法与机器学习法两类。计量指标法依赖论文的引用模式如被引量、颠覆性指数（CD Index）衡量突破性程度，虽然体现了较高的召回率，但也存在难以应对特征关系复杂性、领域适用性不足且准确率较低等问题[11, 17]。相较于计量指标的方案，机器学习模型因在复杂特征关系建模上的卓越能力而在文献计量领域得到了广泛关注，并在高被引论文识别等任务中取得了显著成果[18-19]。而突破性论文识别所涉及的因素更加复杂，相关研究也比较少。代表性研究主要是H. N. Wolcott等[20]将文献的出版时间、共同作者网络及其他字段特征纳入随机森林模型，结合主题专家遴选从而完成识别，准确率和召回率分别达到0.91和0.77。此外，P. Savov等[21]则基于LDA模型和SVM模型根据论文主题分布预测其出版年份，通过与论文实际出版年份的差值度量论文的创新性，从而识别Top 10论文作为潜在突破性论文，但这个方法依赖于论文发表后几年研究主题的发展趋势。整体而言，虽然基于计量指标的方案在一定领域和数据范围内取得了较高的召回率，但领域适用性有限、缺乏建模复杂关系的能力且精准率较低。而基于机器学习的方法虽然能够建模复杂特征关系，但目前在突破性论文识别中相关研究较少且现有研究缺乏对多模型的对比，缺乏对特征重要性的分析，且主要将基于主题创新性作为对创新性的量化。

1.3 创新性类型及其量化指标

创新是对学术论文的本质要求，也是衡量论文学术水

平的重要标准[22]。因此，选择有效的创新性量化方法对于识别突破性论文而言至关重要。T. S. Kuhn[7]提出的科学范式概念将创新性研究划分为颠覆性突破和巩固性渐进，奠定了对学术创新性分类的基础。具体的量化指标方面，相较于从知识组合非典型性[23]、知识单元新颖性[24]等方法对论文主题的创新性进行测度，基于引文知识网络将目标文献置于复杂网络中量化其创新程度。具体来说，R. J . Funk等[25]认为颠覆性创新反映了其在知识网络结构中对既有研究范式的颠覆或巩固程度，定量反映为目标节点前向与后向引证文献在网络中的深层次链接关系，并在此基础上提出了文献网络中颠覆性指数的计算公式，以评估专利对于现有技术趋势的颠覆或巩固程度，为相关研究奠定了基础。随后，Wu等[26]和M. Park等[27]分别将其拓展到科学文献领域。鉴于T. S. Kuhn[7]提出的科学范式概念将创新性研究划分为颠覆性的创新和巩固性的渐进，L. Leydesdorff等[28]、Chen等[29]和Bu等[30]在颠覆性指数的基础上进一步提出颠覆性创新和巩固性创新的具体测度方法，使得能够从上述两个维度对学术论文的创新性进行量化测度。而在最近的研究中，颠覆性和巩固性影响力的思想被应用于各个领域和层次，如颠覆性知识实体的渗透与流动、颠覆性科研成果的演变规律等[4, 27, 31]，体现出其广泛的适用性和解释力，为本文研究创新性指标的选择与构建提供了扎实的理论依据。

2 特征体系构建

首先考虑作者数量、通信作者影响力、通信作者生产力、标题文本长度、文章页数、参考文献数量6个基础特征指标，上述特征指标在高影响力论文识别、学者评估、期刊评估等相关领域得到了广泛应用[18]。由此可以认为，上述特征可能与突破性论文存在一定程度的相关性。其中，作者影响力由h指数表示，因为h指数综合考虑了发文量和被引量，被公认为是评价学者影响力的一种有效指标。作者生产力由作者发表论文数量表示。此外，标题和摘要的文本长度均不包含标点符号。

如表1所示，除基础特征之外，引入了论文发表第N年的被引量、颠覆性指数、知识影响面、颠覆性引文特征、巩固性引文特征等5个特征指标（N的具体取值通过具体实验确定）。其中，颠覆性指数、颠覆性引文特征、巩固性引文特征旨在衡量论文的创新性以及创新类型，以期能够更加准确地识别突破性论文。论文的颠覆性反映了其在知识网络结构中对既有研究范式的颠覆或巩固程度，定量反映为论文的前向与后向引证文献在网络中的深层次链接关系。颠覆性强的论文往往会打破既有研究范式，具备更强的创新价值。R. J. Funk等[25]提出了文献网络中颠覆性指数的计算公式，本文对颠覆性指数的各个参数进行阐释和拓展，将其细分为三个子指标：早期颠覆性被引量、早期巩固性被引量和知识影响面。颠覆性指数是测度科研成果原创性的科学计量指标。如图1所示，在文献网络中，考虑中心论文（fp）及其参考文献（r）的全部被引文献。中心论文颠覆性指数的计算公式为：

$$CD = \frac{N_d - N_c}{N_d + N_c + N_o} \tag{1}$$

式中，N_d、N_c、N_o分别表示只引用fp、同时引用fp和r以及只引用r的文献数量。其中，d反映了论文的颠覆性被引文献[28]，即只引用了中心论文，而没有引用其知识基础（参考文献）的论文，因此N_d可视为颠覆性被引量；c反映了巩固性被引文献[28]是建立在中心论文知识基础上的论文，且中心论文的出现巩固了其在知识网络中的地位，因此N_c可视为巩固性被引量；o反映了中心论文的知识基础的总影响力[16]，反映了中心论文的知识基础在中心论文所在领域之外的影响程度，因此本文将N_o定义为基础知识影响面。

表1 特征体系构建

Tab. 1 Feature system construction

特征编号	特征含义
X_{A1}	作者数量
X_{A2}	通信作者h指数
X_{A3}	通信作者生产力
X_{A4}	标题文本长度
X_{A5}	论文长度
X_{A6}	参考文献数量
X_{B1}	发表后N年被引量
X_{B2}	发表后N年巩固性被引量
X_{B3}	发表后N年颠覆性被引量
X_{B4}	发表后N年知识影响面
X_{B5}	发表后N年颠覆性指数

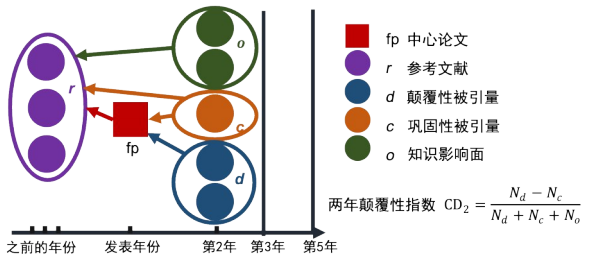


图1 颠覆性被引量、巩固性被引量与知识影响面

Fig. 1 Disruptive citations, consolidating citations, and knowledge impact scope

3 数据和方法

3.1 数据集构建

1）MAG数据集。Microsoft Academic Graph（MAG）

被广泛认为是最大的科学研究文献计量数据库，在规模上超过了Web of Science（WoS）和Scopus。本文使用的MAG数据集有204728474篇海量文献，包括期刊论文、会议论文集和预印本，发表于1800—2021年。其中，77427320篇论文至少被引用过一次，68347900篇论文至少引用过一篇其他论文。这个庞大而全面的数据集为我们提供了丰富的引文数据来源，为准确计算诺奖论文及对照组论文的引文量、颠覆性引用、巩固性引用、参考文献数等指标提供了保障。

2）诺奖论文数据。相较于Zhou等[32]将诺奖得主的所有论文视为突破性论文，Li等[14]通过结合诺奖官网、获奖者个人网站、维基百科条目等不同的信息来源，定位了可能为获得诺奖提供了贡献的论文，从而构建了诺奖论文数据集，以更加精确地区分突破性论文与对照组论文。因此，本文通过Li等[14]提供的诺奖论文ID，在MAG数据集[33]中共确定了712篇发表于1887—2010年物理、化学和医学领域的诺奖论文，从而检验模型在不同领域中的鲁棒性并探索不同领域特征重要性的差异。表2显示了不同领域诺奖论文的数量、平均被引次数、参考文献数、机构数量和平均获奖滞后期。由此可以发现，不同领域诺奖论文的论文数量、平均被引量、平均参考文献数量均存在较大差异。

表2　不同领域诺奖论文描述性统计结果

Tab. 2　Descriptive statistics of Nobel Prize-winning papers in different domains

领域名称	诺奖论文数量/篇	平均被引量/次	平均参考文献数量/篇	平均机构数量/个	平均获奖滞后时间/年
Physics	217	1397	5	1.42	18.2
Chemistry	220	1873	13	1.39	16.0
Medicine	275	1493	12	1.11	16.7

在此基础上，采用Min等[10]相同的方法确定了对照组，即在MAG数据集中获取了与诺奖论文同一年、同一期刊发表的所有论文。在此基础上，获取了所有文献的参考文献和引文，结果如表3所示。

表3　不同领域论文描述性统计结果

Tab. 3　Descriptive statistics of papers in different domains

领域	论文数量/篇	平均被引量/次	平均参考文献数量/篇	平均机构数量/个	诺奖论文占比（%）
Physics	6020	98	2.95	1.42	3.6
Chemistry	9219	99	5.94	1.22	2.38
Medicine	6427	141	8.47	1.18	4.28

3.2　模型选择与介绍

本研究旨在探索突破性论文识别方法，并探索被引量、颠覆性被引量、知识影响面等不同特征变量的重要性。虽然深度学习模型已经在文本、图像领域的各项任务中取得了碾压式的成功，但由于表格型数据（Tabular Data）具有特征不均匀、极值较大等特点，XGBoost[36]等基于树的模型仍然在房价、犯罪预测等大部分任务中取得了目前的最优解[34]。同时，也有研究表明，TabNet[37]和Deep Forest[38]等深度学习模型在部分任务中取得了更好的效果[18, 35]。为此，本文选择了三种具有强解释性的机器学习模型或深度学习模型，以期探索最佳领域突破性论文识别方法并在此基础上探索不同特征的重要性。

XGBoost是一种基于梯度提升树的集成学习模型，通过将多个弱模型的预测结果结合起来产生更准确的预测结果。对于每个节点，XGBoost首先枚举所有特征，然后按照特征值对实例进行排序，并根据信息增益决定不同特征的最佳分裂方案。与其他树集成算法不同，使用预排序算法（Pre-Sorted Algorithm）和基于直方图的算法（Histogram-Based Algorithm）来计算最佳分裂，并通过引入正则化限制模型复杂度从而减少过拟合。

TabNet是一种采用encoder-decoder架构、针对表格型数据的深度神经网络模型，且具有局部可解释性和全局可解释性。其应用于突破性论文识别的优势主要体现在以下三方面：一是通过全连接层、归一化层和非线性激活函数对所选特征进行非线性处理，使得模型可以捕获被引量、颠覆性和巩固性等特征之间复杂的非线性关系；二是利用顺序注意力机制（Sequential Attention）选择最重要的特征，这不仅通过模型重点关注关键特征而提高模型的学习能力，还通过选择最佳的特征路径而使模型具有可解释性；三是与传统的决策树模型相比，TabNet可以直接处理原始表格数据而无须额外预处理，从而减少复杂特征工程的需要。

GcForest是另一种具有强可解释性的深度神经网络模型，同时适用于表格型数据和文本型数据。不同于其他深度学习模型，其在吸收深度模型在逐层处理、内部特征变换以及模型复杂程度方面的优势之外，通过将决策树等不可微模块作为最小单元建立深度模型，在许多任务上表现优异。具体来说，其首先通过多层级联随机森林（Cascade Forest）利用滑动窗口和多粒度扫描（Multi-Grained Scanning）提取多个子样本生成局部特征，通过决策树对局部特征进行初步预测。继而构建多层级联（Cascade of Cascades）进行逐层处理和特征变化，将初步预测结果传递到级联的第一层，用多个随机森林进行训练和预测。将每一层的输出作为下一层的输入，直至达到预定层数或预测精度。

3.3　特征重要性分析方法

准确的特征重要性分析建立在所构建模型准确性高、

稳健性强的基础上[39]。因此，本文在通过对比实验确定各领域最优突破性论文识别模型的前提下，进一步引入了一种模型无关、理论完备的解释方法——SHAP（Shapley Additive Explanations）[40]衡量各类特征指标对于模型预测结果的重要性贡献。

SHAP是一种基于合作博弈论的机器学习模型解释方法，通过计算每个特征加入模型时的边际贡献，以及这个特征在所有特征序列中不同的边界贡献，最后取均值得到SHAP值反映各个特征对模型的影响程度及方向[40]。与依赖于特定模型的方法（如Random Forest中基于平均基尼指数下降的方法）相比，SHAP等方法具有模型无关性，且在处理高基数分类特征或连续特征时不易产生偏差[41]。与同样具有模型无关性的置换重要性（Permutation Importance）方法相比，SHAP允许对特征相关性建模，避免两个特征具有相关性时置换某个特征导致破坏数据结构，从而提升结果的稳定性和准确性。

4 实验和结果

4.1 三种模型识别不同领域突破性论文的效果对比

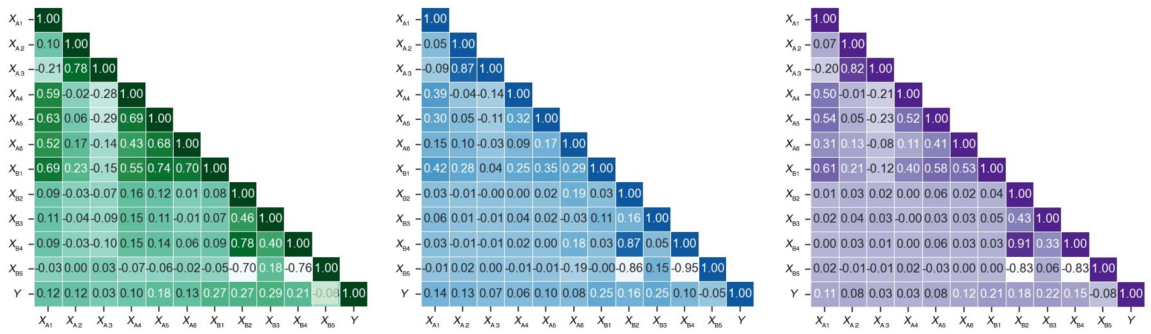
鉴于已有研究表明，1980年之前的论文中部分引文数据的缺失可能导致相关特征变量的计算存在误差[42]。本文选用了1980—2010年的所有诺奖与对照组数据，首先分析了11个特征变量与目标变量之间的相关性，并在此基础上筛选有用特征。相较于皮尔森相关性分析，斯皮尔曼相关性分析（Spearman Correlation Analysis）能够捕捉等级变量（Rank Values）与其他变量之间的相关性，因此本文采用斯皮尔曼相关性分析法。如图2所示，结果表明医学、物理、化学三个领域中X_{B5}（颠覆性指数）与目标变量Y（诺奖论文）之间均为负相关。斯皮尔曼相关性分析结果分别为−0.08、−0.05、−0.08。同时，各学科“X_{B1}”（被引量）、“X_{B2}”（巩固性被引量）、“X_{B3}”（颠覆性被引量）之间的相关性较低，说明在论文发表第10年其巩固性被引与颠覆性被引可能发生了较大分化。基于上述分析，本文在特征体系中剔除了颠覆性指数，后续基于表1中剩余的10个特征对模型进行训练。

在特征选择的基础上，本文通过网格法确定对于每种学科中不同模型的最佳参数，确保每种模型取得最佳效果，每次实验均经过了五折交叉验证以确保研究结果的稳定性。

表4显示了三种不同的机器学习模型在识别1980—2010年物理、化学、医学三个领域的诺奖论文的表现。考虑到诺奖论文在不同领域论文数据集中的占比均较小（2.38%～4.28%），本文分别通过精准率、召回率和F值衡量模型效果。从不同模型的对比来看，XGBoost在数据量、学科均不同的三个数据集上均取得了较好的效果，精确率都达到了90%以上，具有较强的鲁棒性。尽管TabNet在高被引论文识别中取得了较好的效果[10]，但在突破性论文识别任务中，其在三个不同学科中的准确率、召回率和F值均低于XGBoost和GcForest。从不同学科的角度来看，突破性论文识别效果最好的是医学领域，最高精准率和召回率分别达到了97.14%和92.78%，其次是物理领域，最高精确率和召回率分别为97.50%和90.00%。而化学领域突破性论文识别的精确率和召回率均低于其他两个学科。造成这一结果的原因可能与不同学科突破性论文的占比不同有关（医学、物理和化学领域的诺奖论文占比分别为4.28%、3.6%和2.38%）。此外，P. Mitsis[5]的研究表明，化学、医学和物理学领域获奖作品发表后10年内获奖的比例分别为24%、29%和31%，即化学领域的突破性论文往往需要在更长的时间维度中体现其价值。

4.2 引文时间窗口对突破性论文识别效果的影响

对于科技创新的前瞻性布局和科研管理而言，尽早地在论文发表的初期识别突破性论文非常重要。为此，本文进一步对比了分别在论文发表的第10年、第5年和第2年识别不同领域突破性论文的效果，探索引文时间窗口对识别效果的影响，结果如图3、表5所示。



(a) 医学 (b) 物理 (c) 化学

图2 不同领域斯皮尔曼相关性分析结果

Fig. 2 Spearman correlation analysis results across different domains

表4　三种机器学习模型识别突破性论文的效果对比

Tab. 4　Performance comparison of three machine learning models for breakthrough paper identification

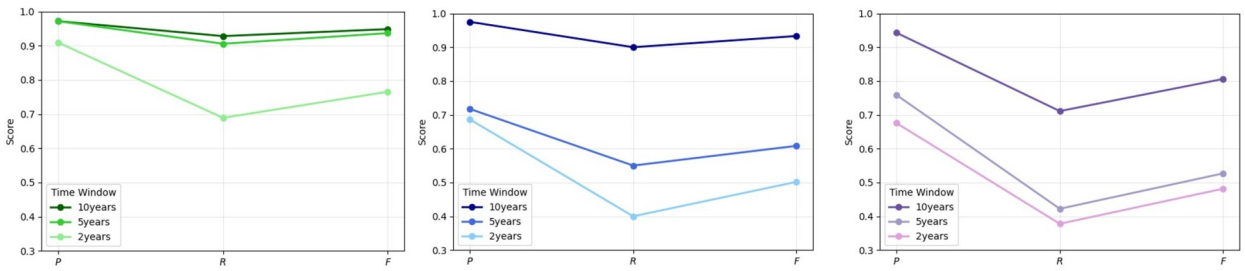
学科	模型	精确率	召回率	*F*值
医学	XGBoost	**0.9714**	**0.9278**	**0.9482**
	TabNet	0.8478	0.7611	0.7843
	GcForest	0.9667	0.9250	0.9429
物理	XGBoost	**0.9750**	**0.9000**	**0.9331**
	TabNet	0.8099	0.6000	0.6840
	GcForest	0.9305	0.9000	0.9111
化学	XGBoost	**0.9428**	**0.7111**	**0.8061**
	TabNet	0.7767	0.5111	0.5922
	GcForest	0.8778	0.6889	0.7528

由表5不难发现，随着时间窗口的缩小，不同领域突破性论文识别的效果均有所下降。其中，物理领域和化学领域的识别效果受时间因素影响较大。物理领域在第5年和第2年的*F*值分别下降了33.21%和41.17%，化学领域则分别下降了25.47%和30.52%。这一结果表明，物理和化学领域难以在论文发表早期识别突破性论文。与之相对应的是，医学领域在发表第5年就能够比较准确地识别突破性论文。这一结果可能与不同学科的引文增长模式存在差异有关[43]。

4.3　领域突破性论文识别中的特征重要性分析

鉴于XGBoost在三个不同领域突破性论文识别中均取得了最好效果，且引文窗口为10年时在不同领域达到了较高的准确率和召回率。本文采用SHAP方法进一步对XGBoost模型中的特征变量进行解释性分析，揭示各特征对突破性论文识别效果的具体贡献。SHAP值越高，表明这个变量对模型输出的影响越显著。

在此前的相关性分析（见图2）中发现，三个不同领



(a) 医学　(b) 物理　(c) 化学

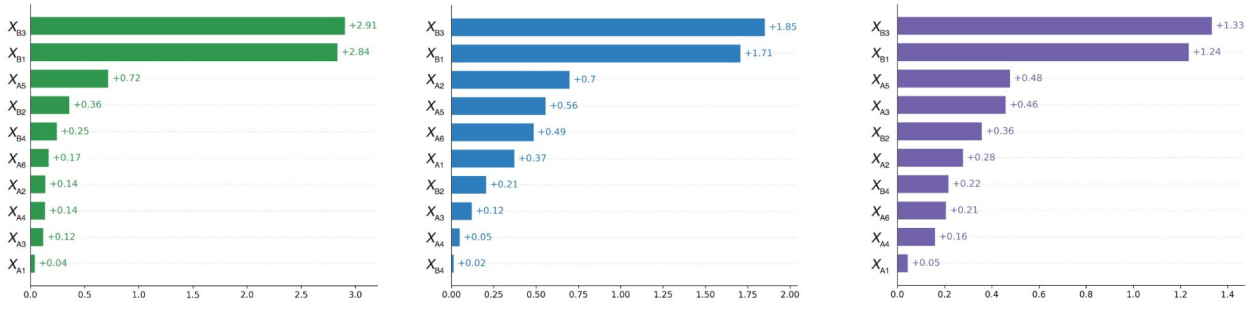
图3　不同引文时间窗口对识别突破性论文的影响

Fig. 3　Impact of different citation time windows on breakthrough paper identification

表5　不同引文时间窗口对识别突破性论文的影响

Tab. 5　Effect of citation time windows on breakthrough paper identification across domains

时间窗口	医学			物理			化学		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
10年	0.9714	0.9278	0.9482	0.9750	0.9000	0.9331	0.9428	0.7111	0.8061
5年	0.9713	0.9055	0.9364	0.7176	0.5500	0.6083	0.7588	0.4222	0.5269
2年	0.9091	0.6889	0.7651	0.6867	0.4000	0.5018	0.6757	0.3778	0.4816



(a) 医学　(b) 物理　(c) 化学

图4　突破性论文识别中指标重要性排序结果

Fig. 4　Feature importance ranking in breakthrough paper identification

域中“X_{B5}”（颠覆性指数）与目标变量“Y”（诺奖论文）之间均为负相关，因此在模型构建阶段已将其剔除。本节分别测量其余10项特征变量在突破性论文识别中的重要性，结果如图4所示。

整体来看，“X_{B3}”（颠覆性引文量）在三个学科中均表现出最高的特征重要性，显著高于引文量等其他特征指标。这表明，颠覆性引文量在不同领域中均是突破性论文成果的重要刻画特征，是精准识别突破性论文的关键因子。“X_{B1}”（被引量）在不同领域的突破性论文识别中同样表现出较高的重要性。尽管已有研究指出，越来越难以直接通过被引量识别诺奖[34]。但本研究表明，被引量仍然是识别突破性论文关键指标之一，尤其是在引用窗口较长时对于突破性论文识别具有较强的贡献。值得注意的是，对于不同学科而言，“X_{B2}”（巩固性引文量）的重要性程度存在差异。总体而言，其对于医学和化学领域的重要性程度高于物理领域，与引文增长模式同样体现出学科差异性[43]，而背后的具体原因需要结合更多不同领域数据及领域专家进一步深入探讨。

5 结束语

本文主要从单篇学术论文评价的角度出发，重点识别其中极具代表性的突破性论文。首先，在作者、参考文献等基础特征的基础上，引入颠覆性和巩固性引文特征，构建了包含11个变量的指标体系。在此基础上，对比了具有可解释性的机器学习XGBoost和两种深度学习模型（TabNet和GcForest）在物理、化学和医学三个不同的学科数据集中识别突破性论文的效果，并选择最优模型进一步探索了特征重要性以及不同引文时间窗口取值对识别效果的影响。研究发现：①在三种不同的模型中，XGBoost在不同领域的突破性论文识别中取得了最佳效果，在医学、物理和化学的$F1$值分别为0.9482、0.9331和0.8061；②对于不同领域而言，模型均在引文窗口为10年时达到最佳识别效果，引文窗口缩小时识别效果显著下降；③颠覆性引文特征与巩固性引文特征能够提升突破性论文识别的效果，巩固性引文特征的具体重要程度对于不同学科而言存在差异。□

参考文献

[1] WINNINK J J, TIJSSEN R J W. Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication [J]. Scientometrics, 2015, 102 (1): 113-134.

[2] LAUDEL G, GLÄSER J. Beyond breakthrough research: epistemic properties of research and their consequences for research funding [J]. Research Policy, 2014, 43 (7): 1204-1216.

[3] PONOMAREV I V, LAWTON B K, WILLIAMS D E, et al. Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? [J]. Scientometrics, 2014, 100 (3): 755-765.

[4] WEI Chunli, LI Jiang, SHI Dongbo. Quantifying revolutionary discoveries: evidence from Nobel prize-winning papers [J]. Information Processing & Management, 2023, 60 (3): 103252.

[5] MITSIS P. The Nobel Prize time gap [J]. Humanities and Social Sciences Communications, 2022, 9 (1): 407.

[6] 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. (ZHOU Haichen, ZHENG Dejun, LI Tianyu. Research on the identification of academic innovation contributions in full academic texts [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39 (8): 845-851.)

[7] KUHN T S. The structure of scientific revolutions [M]. HACKING I. Chicago, IL: University of Chicago Press, 1964.

[8] FUNK R J, OWEN-SMITH J. A dynamic network measure of technological change [J]. Management Science, 2017, 63 (3): 791-817.

[9] YANG A J, HU Haotian, ZHAO Yuehua, et al. From consolidation to disruption: a novel way to measure the impact of scientists and identify laureates [J]. Information Processing & Management, 2023, 60 (5): 103420.

[10] LI Xin, WEN Yang, JIANG Jiaojiao, et al. Identifying potential breakthrough research: a machine learning method using scientific papers and Twitter data [J]. Technological Forecasting and Social Change, 2022, 184: 122042.

[11] 迟培娟，丁洁兰，冷伏海. 突破性论文的三元计量特征及识别研究——以生物医学领域为例 [J]. 情报学报，2022，41 (7)：663-675. (CHI Peijuan, DING Jielan, LENG Fuhai. Characteristics and identification of breakthrough papers: a triple: metric study in the biomedical field [J]. Journal of the China Society for Scientific and Technical Information, 2022, 41 (7): 663-675.)

[12] DU Jian, TANG Xiaoli, WU Yishan. The effects of research level and article type on the differences between citation metrics and F1000 recommendations [J]. Journal of the Association for Information Science and Technology, 2016, 67 (12): 3008-3021.

[13] DU Jian, WU Yishan. A bibliometric framework for identifying “Princes” who wake up the “Sleeping Beauty” in challenge-type scientific discoveries [J]. Journal of Data and Information Science, 2017, 1 (1): 50-68.

[14] LI Jichao, YIN Yian, FORTUNATO S, et al. A dataset of publication records for Nobel laureates [J]. Scientific Data, Nature Publishing Group, 2019, 6 (1): 33.

[15] GINGRAS Y, WALLACE M. Why it has become more difficult to predict Nobel Prize winners: a bibliometric analysis of

Nominees and Winners of the Chemistry and Physics Prizes (1901-2007) [J]. Scientometrics, 2009, 82 (2): 401-412.

[16] INHABER H, PRZEDNOWEK K. Quality of research and the Nobel Prizes [J]. Social Studies of Science, 1976, 6 (1): 33-50.

[17] 杨文霞，孔嘉，闫晓慧，等. 文献颠覆性创新测度——权重颠覆性指数 [J]. 图书情报工作，2024，68 (5)：87-96. (YANG Wenxia, KONG Jia, YAN Xiaohui, et al. Measuring disruptive innovation in literature: a weighted disruption index [J]. Library and Information Service, 2024, 68 (5): 87-96.)

[18] HU Zewen, CUI Jingjing, LIN A. Identifying potentially excellent publications using a citation-based machine learning approach [J]. Information Processing & Management, 2023, 60 (3): 103323.

[19] 张彪，陈云伟，高道斌. 融合学术贡献与创新价值的论文学术影响力预测分级方法 [J]. 图书情报工作，2025，69 (2)：108-120. (ZHANG Biao, CHEN Yunwei, GAO Daobin. Hierarchical method for predicting the academic impact of papers based on the integration of academic contributions and innovative value [J]. Library and Information Service, 2025, 69 (2): 108-120.)

[20] WOLCOTT H N, FOUCH M J, HSU E R, et al. Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers [J]. Scientometrics, 2016, 107 (2): 807-817.

[21] SAVOV P, JATOWT A, NIELEK R, et al. Identifying breakthrough scientific papers [J]. Information Processing & Management, 2020, 57 (2): 102168.

[22] 姜育彦，刘雪立. 绝对颠覆性指数与同行评议指标及CNCI的关系：基于病毒学论文的研究 [J]. 图书情报工作，2023，67 (3)：96-105. (JIANG Yuyan, LIU Xueli. The relationship between absolute disruption index, peer review index and CNCI: a study based on virology papers [J]. Library and Information Service, 2023, 67 (3): 96-105.)

[23] WANG Jian, VEUGELERS R, STEPHAN P. Bias against novelty in science: a cautionary tale for users of bibliometric indicators [J]. Research Policy, 2017, 46 (8): 1416-1436.

[24] CALLAWAY E. Young scientists lead the way on fresh ideas [J]. Nature, 2015, 518 (7539): 283-284.

[25] FUNK R J, OWEN-SMITH J. A Dynamic network measure of technological change [J]. Management Science, 2017, 63 (3): 791-817.

[26] WU Lingfei, WANG Dashun, EVANS J A, et al. Large teams develop and small teams disrupt science and technology [J]. Nature, 2019, 566 (7744): 378-382.

[27] PARK M, LEAHEY E, FUNK R J, et al. Papers and patents are becoming less disruptive over time [J]. Nature, 2023, 613 (7942): 138-144.

[28] LEYDESDORFF L, BORNMANN L. Disruption indices and their calculation using web-of-science data: indicators of historical developments or evolutionary dynamics? [J]. Rochester, NY: Social Science Research Network, 2021, 14 (4): 101219.

[29] CHEN Jiyao, SHAO Diana, FAN Shaokun, et al. Destabilization and consolidation: conceptualizing, measuring, and validating the dual characteristics of technology [J]. Research Policy, 2021, 50 (1): 104115.

[30] BU Yi, WALTMAN L, HUANG Yong, et al. A multidimensional framework for characterizing the citation impact of scientific publications [J]. Quantitative Science Studies, 2021, 2 (1): 155-183.

[31] 刘小慧，朱曼曼. 科研论文颠覆性与参考文献年龄之间的关系研究 [J]. 情报理论与实践，2023，46 (7)：60-66. (LIU Xiaohui, ZHU Manman. Research on the relationship between the disruption of scientific papers and the age of references [J]. Information Studies: Thery & Application, 2023, 46 (7): 60-66.)

[32] ZHOU Zhiwei, XING Rui, LIU Jing, et al. Landmark papers written by the Nobelists in physics from 1901 to 2012: a bibliometric analysis of their citations and journals [J]. Scientometrics, 2014, 100 (2): 329-338.

[33] WANG Kuansan, SHEN Zhihong, HUANG Chiyuan, et al. Microsoft academic graph: when experts are not enough [J]. Quantitative Science Studies, 2020, 1 (1): 396-413.

[34] GRINSZTAJN L, OYALLON E, VAROQUAUX G. Why do tree-based models still outperform deep learning on typical tabular data? [C]//Advances in Neural Information Processing Systems. New Orleans, LA: Curran Associates Inc., 2022: 507-520.

[35] ARNOULD L, BOYER C, SCORNET E. Analyzing the tree-layer structure of deep forests [C]//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 2021: 342-350.

[36] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery, 2016: 785-794.

[37] ARIK S Ö, PFISTER T. TabNet: attentive interpretable tabular learning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35 (8): 6679-6687.

[38] ZHOU Zhihua, FENG Ji. Deep Forest: towards an alternative to deep neural networks [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017: 3553-3559.

[39] RUDIN C, CHEN Chaofan, CHEN Zhi, et al. Interpretable machine learning: fundamental principles and 10 grand challenges [J]. Statistic Surveys, 2022 (6): 1-85.

（下转第81页）

ing methods of opinion leaders in health super topic and its guiding effect on different participation behaviors［J］. Library and Information Service，2023，67（4）：91-101.）

［47］裘江南，徐雪冬，谷文静，等. 突发公共卫生事件下不同人格特质公众的信息需求比较分析［J］. 数据分析与知识发现，2024，8（6）：119-131.（QIU Jiangnan，XU Xuedong，GU Wenjing，et al. Comparing information needs of users with different personalities in public health emergencies［J］. Data Analysis and Knowledge Discovery，2024，8（6）：119-131.）

［48］喻雪寒，何琳，徐健. 基于RoBERTa-CRF的古文历史事件抽取方法研究［J］. 数据分析与知识发现，2021，5（7）：26-35.（YU Xuehan，HE Lin，XU Jian. Extracting events from ancient books based on RoBERTa-CRF［J］. Data Analysis and Knowledge Discovery，2021，5（7）：26-35.）

［49］张凯亮，臧国全，肖洋. 医疗信息文本中的个人隐私数据识别与计量研究［J］. 情报学报，2024，43（8）：936-945.（ZHANG Kailiang，ZANG Guoquan，XIAO Yang. Research on privacy data identification and measurement based on medical information text［J］. Journal of the China Society for Scientific and Technical Information，2024，43（8）：936-945.）

［50］ZHOU Lexin，SCHELLAERT W，MARTÍNEZ-PLUMED F，et al. Larger and more instructable language models become less reliable［J］. Nature，2024：1-8.

［51］PARK C L. Making sense of the meaning literature：an integrative review of meaning making and its effects on adjustment to stressful life events［J］. Psychological Bulletin，2010，136（2）：257-301.

［52］NICOL E，BRIGGS J，MONCUR W，et al. Revealing cumulative risks in online personal information：a data narrative study［J］. Proceedings of the ACM on Hu-man-Computer Interaction，2022，6（CSCW2）：1-25.

［53］梅潇，查先进，严亚兰. 智能推荐环境下移动社交媒体用户隐私风险感知影响机理研究［J］. 情报理论与实践，2024，47（1）：57-64.（MEI Xiao，ZHA Xianjin，YAN Yalan. Influencing mechanism of privacy risk perception in the context of mobile social media intelligent recommendation［J］. Information Studies：Theory & Application，2024，47（1）：57-64.）

［54］宋姝. 房地产登记数据资源开发利用的法律规制［J］. 政治与法律，2024（3）：148-159.（SONG Shu. On legal regulation of exploitation and utilization of real estate registration data resources［J］. Political Science and Law，2024（3）：148-159.）

［55］国家市场监督管理总局，国家标准化管理委员会. 基于云计算的重大突发公众卫生事件中社区服务系统基本要求：GB/T44229—2024［S/OL］.［2025-06-14］. http：//c. gb688. cn/bzgk/gb/showGb？type=online&hcno=0869F865A355D44822F43CFECAA4B80D.

作者简介：耿瑞利，女，博士，副教授。研究方向：隐私管理，数据治理，社交媒体用户信息行为。**李森涛**（通信作者，Email：lisentao99@163.com），男，博士生。研究方向：数据隐私与数据治理。**魏琦**，女，硕士生。研究方向：数据隐私与数据治理。

作者贡献声明：耿瑞利，论文选题，论文修改与定稿。**李森涛**，研究设计，数据采集与分析，论文撰写。**魏琦**，论文修改。

录用日期：2025-09-22

（上接第133页）

［40］LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions［C］//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook，NY：Curran Associates Inc.，2017：4768-4777.

［41］MAHMUD S，REDOWAN M，AHMED R，et al. Phenology-based classification of Sentinel-2 data to detect coastal mangroves［J］. Geocarto International，2022，37（26）：14335-14354.

［42］LIANG Guoqiang，LOU Ying，HOU Haiyan. Revisiting the disruptive index：evidence from the Nobel Prize-winning articles［J］. Scientometrics，2022，127（10）：5721-5730.

［43］AKSNES D W，AKSNES D W，AKSNES D W，et al. Characteristics of highly cited papers［J］. Research Evaluation，2003，12（3）：159-170.

作者简介：张琪，女，1995年生，博士，讲师。研究方向：信息计量与科学评价，数字人文。**马瑞敏**，男，1983年生，博士，教授，博士生导师。研究方向：信息计量与科学评价，数字人文。**杨杰**，博士生。研究方向：科学计量，复杂网络，创新测度。**王左戎**，女，1998年生，博士生。研究方向：信息组织与检索，科学计量。**邓三鸿**（通信作者，Email：sanhong@nju.edu.cn），男，1975年生，博士，教授，博士生导师。研究方向：领域知识图谱，科学计量，信息分析与科学评价。

作者贡献声明：张琪，论文框架，模型训练，论文撰写。**马瑞敏**，论文框架，修改定稿。**杨杰**，数据整理，论文撰写。**王左戎**，数据分析。**邓三鸿**，提出论文思路，修订论文。

录用日期：2025-09-04